3 SEMICONDUCTOR SCIENCE AND LIGHT-EMITTING DIODES

3.1 REVIEW OF SEMICONDUCTOR CONCEPTS AND ENERGY BANDS

A. Energy Band Diagrams, Density of States, Fermi–Dirac Function and Metals

We know from modern physics that the energy of the electron in an atom is quantized and can only have certain discrete values as illustrated for the Li atom in Figure 3.1. The isolated Li atom has two electrons in the 1*s* shell and one electron in the 2*s* subshell. The same concept also applies to the electron energy in a molecule with several atoms. Again the electron energy is quantized. However, when we bring together something like 10^{23} Li atoms to form the metal crystal, the interatomic interactions result in the formation of electron **energy bands**. The 2*s* energy level splits into some 10^{23} closely spaced energy levels that effectively form an **energy band**, which is called the 2*s* band. Similarly other higher energy levels also form bands as illustrated in Figure 3.1. These energy bands overlap to form one continuous energy band which represents the energy band structure of a metal. The 2*s* energy level in the Li atom is half full (2*s* subshell needs 2 electrons but has 1) which means that the 2*s* band in the crystal will also be half full. Metals characteristically have partially filled energy bands.

In the case of metals, the energy band diagram is straightforward and essentially one band of energies in which the energy levels from the bottom of the band to a certain level are all filled at zero Kelvin (0 K), as indicated in Figure 3.1. Each energy level in this band corresponds to a wavefunction for the electron inside the crystal, and represents a **quantum state** for the electron. Thus, the energy levels are often referred to as energy states. In the solid, the bottom of the band is normally taken as the zero of energy so that electrons fill all the states until a certain energy called the **Fermi energy**, denoted as E_F—the highest occupied energy level at 0 K. The states above E_F are empty up to the vacuum level. The energy required to remove an electron from the metal is the energy involved in taking an electron from E_F to the vacuum level, which is called the **work function** Φ of the metal. For example, a photon with an energy $h\upsilon$ greater than Φ that is incident on a metal can cause an electron at E_F to become ejected out from the metal to become free, a phenomenon called the **photoelectric effect**. The electron is said to be **photoemitted.** Many photomultiplier tubes in optoelectronics have a cathode material with the right Φ that allows electrons to be photoemitted when light within the desired range of wavelengths is incident on the cathode.

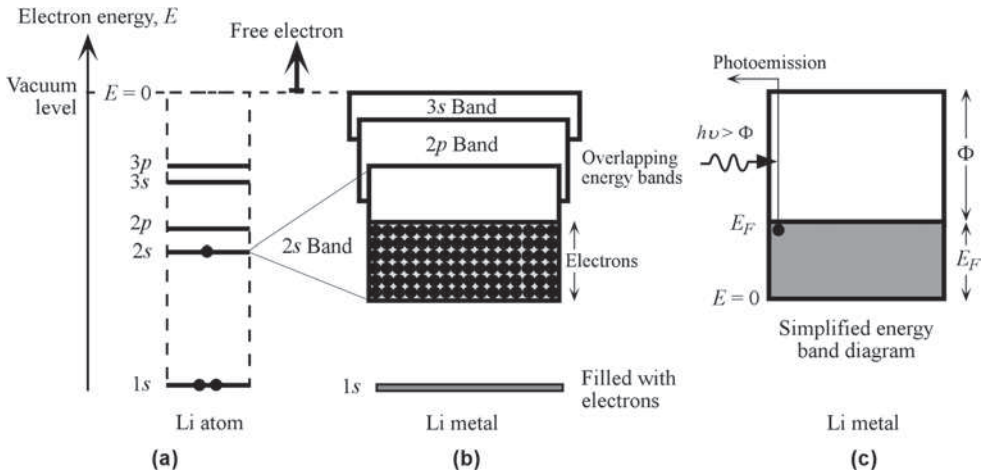


FIGURE 3.1 (a) Energy levels in a Li atom are discrete. (b) The energy levels corresponding to outer shells of isolated Li atoms form an energy band inside the crystal—for example, the 2*s* level forms a 2*s* band. Energy levels form a quasi-continuum of energy within the energy band. Various energy bands overlap to give a single band of energies that is only partially full of electrons. There are states with energies up to the vacuum level where the electron is free. (c) A simplified energy band diagram for a metal, and the photoelectric effect.

The photoemitted electrons are then accelerated by the applied field inside the multiplier tube and are made incident on other electrodes to cause secondary electron emission and hence multiplication. (One incident photon can give rise to 10^6 collected electrons.)

The electrons around E_F control many of the properties of metals such as electrical and thermal conductivity. When an electric field E is applied, an electron at E_F can gain energy from the external force eE and be able to move to a higher energy state since these are empty and available. Electrical conduction occurs by the drift of electrons at around E_F.

Many important properties of both metals (and semiconductors below) are described by considering the behavior of electrons within an energy band. In the case of metals, this is called the **free electron model**; that is, the electrons are free inside the metal and hence electron potential energy is constant. These electrons are what are normally called conduction electrons.

There are two important concepts in describing free electrons in an energy band. The **density of states** (DOS) $g(E)$ represents the number of electronic states (electron wavefunctions) in a band per unit energy per unit volume of the crystal. We can use quantum mechanics to calculate the DOS by considering how many electron wavefunctions there are within a given energy range per unit volume of the crystal. Figure 3.2 (a) shows the energy band diagram of a metal, and (b) shows, in a simplified way, how $g(E)$ depends on the electron energy in the band. According to quantum mechanics, for an electron confined within a three-dimensional potential energy well, as one might expect for a conduction electron in the crystal of a metal, DOS $g(E)$ increases with energy from the bottom of the band and is given by

Density of states in a band

$$g(E) = 4\pi (2m_e)^{3/2} h^{-3} E^{1/2} = AE^{1/2} \qquad \textbf{(3.1.1)}$$

where E is the electron energy from the bottom of the band, m_e is the mass of the electron, h is Planck's constant, and A is a constant that represents the quantities multiplying $E^{1/2}$ in the second term. The DOS gives information on available states, that is, wavefunctions, for the electron and not on their actual occupation.

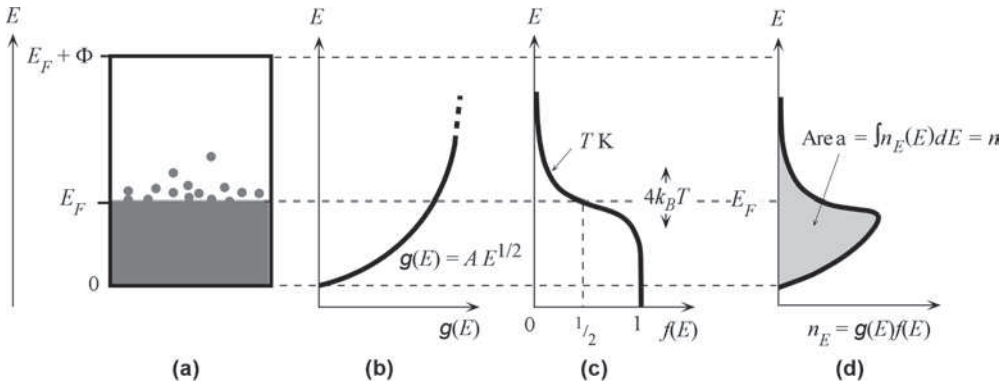


FIGURE 3.2 (a) Above 0 K, due to thermal excitation, some of the electrons are at energies above E_F. (b) The density of states, $g(E)$ vs. E in the band. (c) The probability of occupancy of a state at an energy E is $f(E)$. The product $g(E)f(E)$ is the number of electrons per unit energy per unit volume or electron concentration per unit energy. The area under the curve with the energy axis is the concentration of electrons in the band, n.

The **Fermi–Dirac function** $f(E)$ is the probability of finding an electron in a quantum state with energy E (state implies a wavefunction). It is a fundamental property of a collection of interacting electrons in *thermal equilibrium.* It is given by

$$f(E) = \frac{1}{1 + \exp\left(\dfrac{E - E_F}{k_B T}\right)} \tag{3.1.2}$$

Fermi–Dirac function

where k_B is the Boltzmann constant, T is the temperature (K), and E_F is the Fermi energy. Figure 3.2 (c) shows the dependence of $f(E)$ on the energy E. At low energies well below E_F, all states are occupied and at high energies, well above E_F, the states are empty. At E_F, $f(E_F) = 1/2$, half the states are occupied. Changes in f(E) occur around E_F over several k_BTs of energy, as indicated in Figure 3.2 (c). The higher the temperature, the greater is the reach of $f(E)$ to higher energies, and therefore more electrons at higher energies.

Consider the product $g(E)f(E)$, density of states at $E \times$ probability of finding an electron in a state at E. The product $g(E)f(E)$ represents the number electrons per unit energy per unit volume, $n_E(E)$, in the band which is shown in Figure 3.2 (d). It is essentially the energy distribution of electrons in the band. Thus, $n_E\,dE = g(E)f(E)dE$ is the number of electrons in the energy range E to $E + dE$. Integrating this from the bottom to the top of the band gives the conduction electron concentration n in the band,

$$n = \int_0^{E_F+\Phi} g(E)f(E)dE \tag{3.1.3}$$

Equations (3.1.1) and (3.1.2) can be substituted in Eq. (3.1.3) to evaluate the integration. Once integrated, the resulting equation will relate n to the Fermi energy E_F at a temperature T. Put differently, E_F depends on n, the total number of electrons per unit volume in the band. At 0 K, the Fermi energy E_{FO} is

$$E_{FO} = \left(\frac{h^2}{8m_e}\right)\left(\frac{3n}{\pi}\right)^{2/3} \tag{3.1.4}$$

Fermi energy at 0 K for metals

E_{FO} is typically a few electron volts, that is, orders of magnitude greater than k_BT. Above 0 K, E_F decreases slightly with T, though we ignore this tiny change in using metal band diagrams in optoelectronics. Inasmuch as the electrons in the metal are free, they have kinetic energies only. An electron at the Fermi energy E_{FO} has several eVs of kinetic energy, which is orders of magnitude greater than the classical kinetic theory (roughly $3k_BT/2$ for a free particle). The reason is that the electrons in the metal must obey the *Pauli exclusion principle*, which prevents more than two electrons (with spins up and down) from occupying a given energy state. Thus, the electrons must distribute themselves among available states and need to reach higher energies to satisfy the Pauli exclusion principle.

B. Energy Band Diagrams of Semiconductors

The electron energies in a semiconductor crystal, however, are distinctly different than that for metals. Figure 3.3 (a) shows a simplified two-dimensional view of the silicon crystal which has each Si atom bonding to four neighbors. All the four valence electrons per atom are used in these bonds. The interactions between the Si atoms and their valence electrons result in the electron energy in the crystal falling into two distinct energy bands called the **valence band** (VB) and **conduction band** (CB) that are separated by an energy gap, **bandgap** (E_g), as shown in Figure 3.3 (b). There are no allowed electron energies in the **bandgap**; it represents the forbidden electron energies in the crystal. The valence band represents electron wavefunctions in the crystal that correspond to bonds between the atoms. Electrons that occupy these wavefunctions are the valence electrons. Since at absolute zero of temperature all the bonds are occupied by valence electrons (there are no broken bonds), all the energy levels in the VB are normally filled with these electrons. The CB represents electron wavefunctions in the crystal that have higher energies than those in the VB and are normally empty at zero Kelvin. The top of the VB is labeled E_v, bottom of conduction band E_c, so that $E_g = E_c - E_v$ is the bandgap. The width of the CB is called the **electron affinity** χ.

An electron placed in the CB is free to move around the crystal and also to respond to an electric field because there are plenty of neighboring empty energy levels. This electron can easily gain energy from the field and move to higher energy levels because these states are

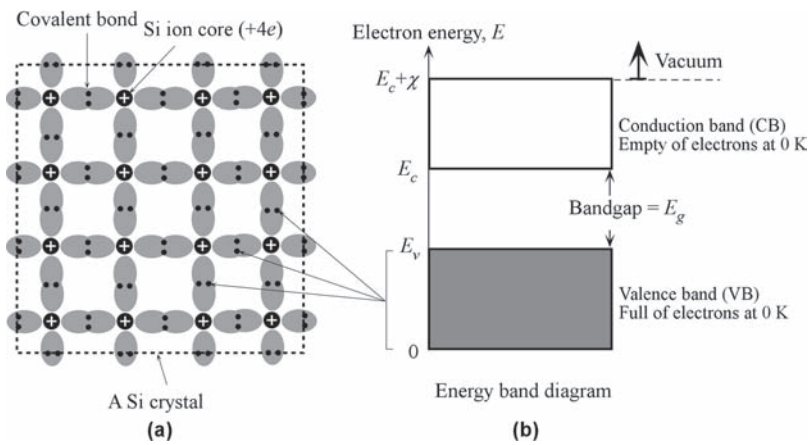


FIGURE 3.3 (a) A simplified two-dimensional view of a region of the Si crystal showing covalent bonds. (b) The energy band diagram of electrons in the Si crystal at absolute zero of temperature. The bottom of the VB has been assigned a zero of energy.

empty in the CB. Generally we can treat an electron in the CB as if it were free within the crystal by simply assigning an **effective mass** m_e^* to it. This effective mass is a quantum mechanical quantity, which takes into account that the electron in the CB interacts with a periodic potential energy as it moves through the crystal so that its inertial resistance to acceleration (definition of mass) is not the same as if it were free in vacuum.

Since the only empty states are in the CB, the excitation of an electron from the VB requires a minimum energy of E_g. Figure 3.4 illustrates what happens when an incident photon of energy $h\upsilon > E_g$ interacts with an electron in the VB. This electron absorbs the incident photon and gains sufficient energy to surmount the energy gap E_g and reach the CB. Consequently, a free electron in the CB and a "hole," corresponding to a missing electron in the VB, are created. In some semiconductors, such as Si and Ge, the photon absorption process also involves lattice vibrations (vibrations of the Si atoms) which we have not shown in Figure 3.4.

The empty electronic state, or the missing electron in the bond, is what we call a **hole** in the valence band. The free electron, which is in the CB, can wander around the crystal and contribute to the electrical conduction when an electric field is applied. The region remaining around the hole in the VB is positively charged because a charge of $-e$ has been removed from an otherwise neutral region of the crystal. This hole, denoted as h^+, can also wander around the crystal as if it were "free." This is because an electron in a neighboring bond can "jump," that is, tunnel into the hole, to fill the vacant electronic state at this site and thereby create a hole at its original position as illustrated in Figure 3.5. This is effectively equivalent to the hole being displaced in the opposite direction. Thus, conduction in semiconductors occurs by both electrons and holes with charges $-e$ and $+e$ and their own effective masses m_e^* and m_h^*.

When a wandering electron in the CB meets a hole in the VB, it has found an empty electronic state of lower energy and it therefore occupies it. The electron falls from the CB to the VB to fill the hole. This is called **recombination**, which results in the annihilation of an electron in the CB and a hole in the VB as in Figure 3.5. The excess energy of the electron falling from CB to VB in certain semiconductors such as GaAs and InP is emitted as a photon. In Si and Ge the excess energy is lost as lattice vibrations (heat), that is, as phonons. In the steady state, the thermal generation rate is balanced by the recombination rate so that the electron concentration n in the CB and hole concentration p in the VB remain constant; both n and p depend on the temperature.

Although in the specific example in Figure 3.4, a photon of energy $h\upsilon > E_g$ creates an electron–hole pair, other sources of energy can also lead to an electron–hole pair creation. In fact, in the absence of radiation, there is still an electron and hole pair generation process going on in the sample as a result of **thermal generation**. Due to thermal energy, the atoms in the crystal

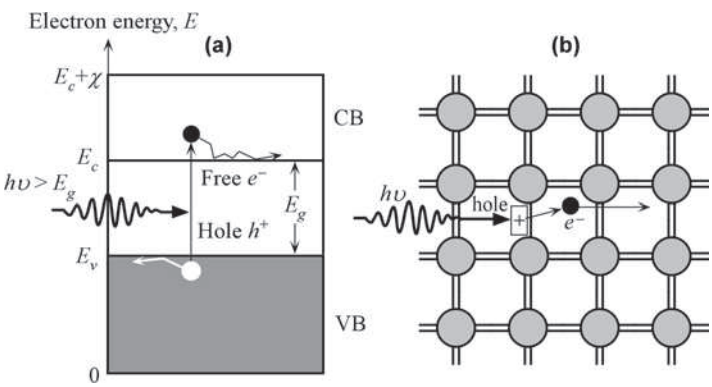


FIGURE 3.4 (a) A photon with an energy $h\upsilon$ greater than E_g can excite an electron from the VB to the CB. (b) Each line with Si-Si atoms is a valence electron in a bond. When a photon breaks a Si-Si bond, a free electron and a hole in the Si-Si bond are created. The result is the photogeneration of an electron–hole pair (EHP).

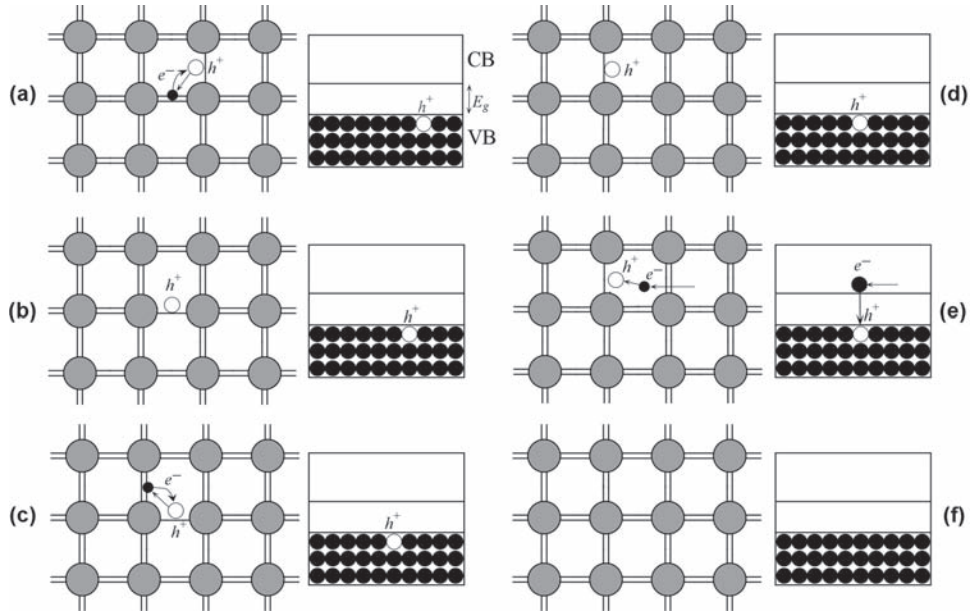


FIGURE 3.5 A pictorial illustration of a hole in the valence band wandering around the crystal due to the tunneling of electrons from neighboring bonds, and its eventual recombination with a wandering electron in the conduction band. A missing electron in a bond represents a hole as in (a). An electron in a neighboring bond can tunnel into this empty state and thereby cause the hole to be displaced as in (a) to (d). The hole is able to wander around in the crystal as if it were free but with a different effective mass than the electron. A wandering electron in the CB meets a hole in the VB in (e), which results in the recombination and the filling of the empty VB state as in (f).

are constantly vibrating which corresponds to the bonds between the Si atoms being periodically deformed with a distribution of energies. Energetic vibrations can rupture bonds and thereby create electron–hole pairs, which correspond to the excitation of electrons from the VB to the CB.

3.2 SEMICONDUCTOR STATISTICS

Many important properties of semiconductors are described by considering the behavior of electrons in the CB and holes in the VB. We need to consider the density of states (DOS) $g_{CB}(E)$ of electronic states in the CB, $g_{VB}(E)$ in the VB and their occupation statistics. Figure 3.6 (a) shows the energy band diagram of a semiconductor, and (b) shows, in a simplified way, how $g_{CB}(E)$ depends on the electron energy in the CB and VB near the band edges E_c and E_v, respectively. The DOS $g_{CB}(E)$ increases with energy from the CB edge as

Density of states near E_c

$$g_{CB}(E) = 4\pi(2m_e^*)^{3/2}h^{-3}(E - E_c)^{1/2} \quad \textbf{(3.2.1)}$$

where $(E - E_c)$ is the electron energy from the bottom of the CB. The DOS curves usually have various peaks at certain energies for crystalline solids, which we have neglected as these are not needed for the present discussion.

To find the energy distribution of electrons in the CB, we need the Fermi–Dirac function $f(E)$, the probability of finding an electron in a quantum state with energy E. The behavior of $f(E)$ is shown in Figure 3.6 (c) assuming that the Fermi level E_F is located in the bandgap. (Its position

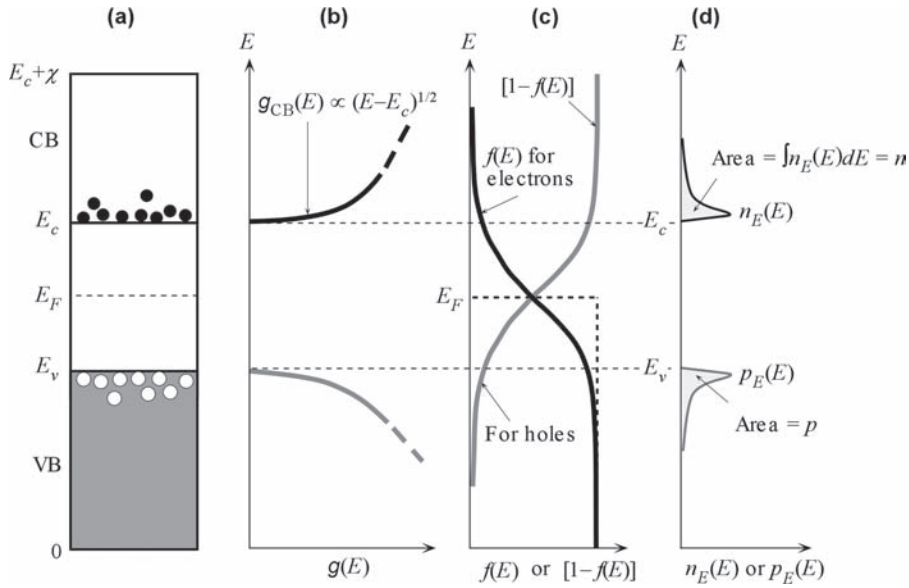


FIGURE 3.6 (a) Energy band diagram of a semiconductor, assuming that the Fermi level is near the middle of the bandgap. (b) Density of states (number of states per unit energy per unit volume). (c) Fermi–Dirac probability function (probability of occupancy of a state). (d) The product of $g(E)$ and $f(E)$ is the energy density of electrons in the CB (number of electrons per unit energy per unit volume). The area under $n_E(E)$ vs. E is the electron concentration.

in the energy gap is discussed later.) When we multiply $g(E)$ with $f(E)$, we get the number of electrons per unit energy per unit volume at E in the CB, $n_E(E)$, which is shown in Figure 3.6 (d). Thus, $n_E dE = g_{CB}(E)f(E)dE$ is the number of electrons in the energy range [Figure 3.6 (d)] E to $E + dE$. Integrating this from the bottom (E_c) to the top ($E_c + \chi$) of the CB gives the electron concentration

$$n = \int_{E_c}^{E_c+\chi} g_{CB}(E)\, f(E)\, dE \qquad \textbf{(3.2.2)}$$

Electron concentration in CB

Whenever $(E_c - E_F) \gg k_B T$, that is, E_F is at least a few $k_B T$ below E_c, then $f(E) \approx \exp\left[-(E - E_F)/k_B T\right]$. That is, the Fermi–Dirac statistics can be replaced by *Boltzmann* statistics. Such semiconductors are called **nondegenerate**. It implies that the number of electrons in the CB is far less than the number states in this band. For nondegenerate semiconductors, the above integration leads to

$$n = N_c \exp\left(-\frac{E_c - E_F}{k_B T}\right) \qquad \textbf{(3.2.3)}$$

Electron concentration in CB

where $N_c = 2\left[2\pi m_e^* k_B T/h^2\right]^{3/2}$ is a constant at a given temperature for a given material, called the **effective density of states at the CB edge**. The result of the integration in Eq. (3.2.3) seems too simple but it is, however, an approximation as it assumes that $(E_c - E_F) \gg k_B T$. We can interpret Eq. (3.2.3) as follows. If we take all the states in the conduction band and replace them with an effective concentration N_c (number of states per unit volume) at E_c and then multiply this simply by the Boltzmann probability function, $f(E_c) = \exp\left[-(E_c - E_F)/k_B T\right]$, we would

obtain the concentration of electrons at E_c, that is, in the conduction band. N_c is thus an effective density of states at the CB band edge.

We can carry out a similar analysis for the concentration of holes in the VB as illustrated in Figures 3.6 (a)–(d). Multiplying the density of states $g_{\text{VB}}(E)$ in the VB with the probability of occupancy by a hole, $\left[1 - f(E)\right]$, gives p_E, the hole concentration per unit energy. Note that the probability of finding a hole, a missing electron, in a state with energy E is $1 - f(E)$. Integrating $p_E(E)$ over the VB gives the hole concentration by assuming that E_F is a few k_BT above E_v, we obtain

Hole concentration in VB

$$p \approx N_v \exp\left(-\frac{E_F - E_v}{k_BT}\right) \tag{3.2.4}$$

where $N_v = 2\left[2\pi m_h^* k_B T/h^2\right]^{3/2}$ is the **effective density of states at the VB edge**, and m_h^* is the effective mass of holes in the VB.

There are no specific assumptions in our derivations above, except for E_F being a few k_BT away from the band edges, which means that Eqs. (3.2.3) and (3.2.4) are generally valid. It is apparent from Eqs. (3.2.3) and (3.2.4) that the location of E_F determines the electron and hole concentrations. Thus, E_F can be viewed as a useful semiconductor property that allows us to represent electron and hole concentrations in the energy band diagram. In an intrinsic semiconductor (a pure crystal), $n = p$, by using Eqs. (3.2.3) and (3.2.4) we can show that the Fermi level E_{Fi} in the *intrinsic crystal* is above E_v and located in the bandgap at

Fermi level in intrinsic crystal

$$E_{Fi} = E_v + \tfrac{1}{2} E_g - \tfrac{1}{2} kT \ln\left(\frac{N_c}{N_v}\right) \tag{3.2.5}$$

Typically N_c and N_v values are comparable and both occur in the logarithmic term so that E_{Fi} is approximately in the middle of the bandgap as originally sketched in Figure 3.6.

There is a useful semiconductor relation between n and p, called the **mass action law**. From Eqs. (3.2.3) and (3.2.4), the product np is

Mass action law

$$np = N_c N_v \exp\left(-\frac{E_g}{k_BT}\right) = n_i^2 \tag{3.2.6}$$

where $E_g = E_c - E_v$ is the bandgap energy and n_i^2 has been *defined as* $N_c N_v \exp(-E_g/k_BT)$. The quantity n_i is a material property; for example, it depends on E_g, but not on the position of the Fermi level. It also depends on the temperature. The intrinsic concentration n_i corresponds to the concentration of electrons or holes in an undoped (pure) crystal, that is, intrinsic semiconductor. In such a semiconductor $n = p = n_i$, which is therefore called the **intrinsic concentration**. The mass action law is valid whenever we have thermal equilibrium and the sample is in the dark (without any photogeneration).

Equations (3.2.3) and (3.2.4) determine the total concentration of electrons and holes in the CB and VB, respectively. The average energy of the electrons in the CB can be calculated by using $n_E(E)$, their energy distribution. The result gives the average energy as $(3/2)k_BT$ above E_c. Since the electron in the CB is "free" in the crystal with an effective mass m_e^*, it wanders around the crystal with an average kinetic energy $(3/2)k_BT$, which is the same as the kinetic energy of a free atom in a gas or vapor in a tank. This is not surprising as both particles are wandering freely (without interacting with each other) and obey Boltzmann statistics. If v is the electron velocity and triangular brackets represent an average, then $<(1/2)m_e^* v^2>$ must be $(3/2)k_BT$. We can thus calculate the *root mean square velocity* $\left(\langle v^2 \rangle\right)^{1/2}$ which is called the **thermal velocity** (v_{th}) and is typically $\sim 10^5$ m s^{-1}. The same ideas apply to holes in the VB with an effective hole mass m_h^*.

It is important to understand what the **Fermi energy** represents. First, it is a convenient way to represent free carrier concentrations (n in the CB and p in the VB) on the energy band diagram. However, the most useful property of E_F is in terms of a change in E_F. Any change ΔE_F across a material system represents electrical work input or output per electron.[2] If V is the potential difference between two points in a material system, then

Fermi energy and electrical work

$$\Delta E_F = eV \tag{3.2.7}$$

For a semiconductor system in equilibrium, in the dark, and with no applied voltage or no EMF generated, $\Delta E_F = 0$ and E_F must be uniform across the system.

3.3 EXTRINSIC SEMICONDUCTORS

A. *n*-Type and *p*-Type Semiconductors

By introducing small amounts of impurities into an otherwise-pure crystal, it is possible to obtain a semiconductor in which the concentration of carriers of one polarity is much in excess of the other type. Such semiconductors are referred to as **extrinsic semiconductors** vis-à-vis the intrinsic case of a pure and perfect crystal. For example, by adding pentavalent impurities, such as arsenic, which have a valency one more than Si, we can obtain a semiconductor in which the electron concentration is much larger than the hole concentration. In this case we will have an ***n*-type semiconductor**. If we add trivalent impurities, such as boron, which have a valency of one less than four, we then have an excess of holes over electrons—a ***p*-type semiconductor**.

An arsenic (As) atom has five valence electrons whereas Si has four. When the Si crystal is doped with small amounts of As, each As atom substitutes for one Si atom and is surrounded by four Si atoms. When an As atom bonds with four Si atoms, it has one electron left unbounded. This fifth electron cannot find a bond to go into so it is left orbiting around the As atom, which looks like an As^+, as illustrated in Figure 3.7 (a). The As^+ ionic center, with an electron e^- orbiting

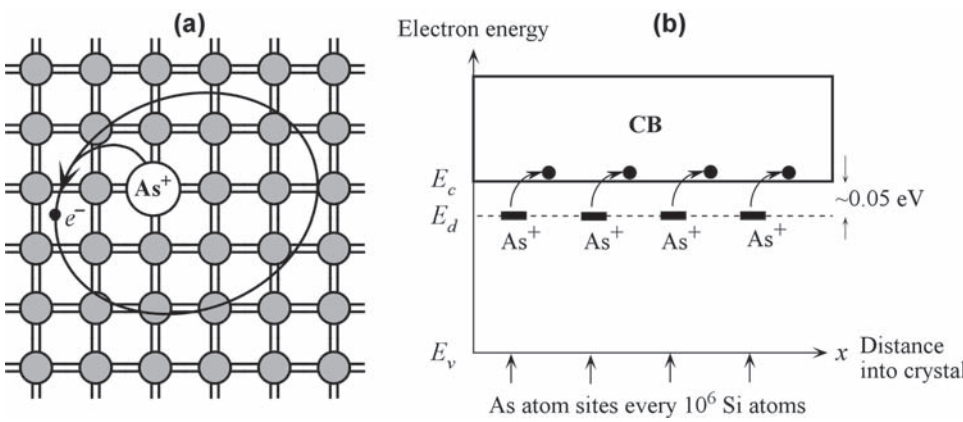


FIGURE 3.7 (a) The four valence electrons of As allow it to bond just like Si but the fifth electron is left orbiting the As site. The energy required to release to free fifth electron into the CB is very small. (b) Energy band diagram for an n-type Si doped with 1 ppm As. There are donor energy levels just below E_c around As^+ sites.

[2]For readers familiar with thermodynamics, its rigorous definition is that E_F is the *chemical potential per electron*, that is, Gibbs free energy per electron. The definition in Eqs. (3.2.7) is in terms of a change in E_F.

it, resembles a hydrogen atom in a silicon environment. We can easily calculate how much energy is required to free this electron away from the As site, thereby ionizing the As impurity by using our knowledge on the ionization of a hydrogen atom (removing the electron from the H-atom). This energy turns out to be a few hundredths of an electronvolt, that is, ~0.05 eV, which is comparable to the thermal energy at room temperature ($\sim k_B T = 0.025$ eV). Thus, the fifth valence electron can be readily freed by thermal vibrations of the Si lattice. The electron will then be "free" in the semiconductor, or in other words, it will be in the CB. The energy required to excite the electron to the CB is therefore ~0.05 eV. The addition of As atoms introduces localized electronic states at the As sites because the fifth electron has a localized wavefunction, of the hydrogenic type, around As^+. The energy of these states, E_d, is ~0.05 eV below E_c because this is how much energy is required to take the electron away into the CB. Thermal excitation by lattice vibrations at room temperature is sufficient to ionize the As atom, that is, excite the electron from E_d into the CB. This process creates free electrons; however, the As^+ ions remain immobile as shown in the energy band diagram of an *n*-type semiconductor in Figure 3.7 (b).

Because the As atom donates an electron into the CB, it is called a **donor** impurity. E_d is the electron energy around the donor atom and it is below E_c by ~0.05 eV as in Figure 3.7 (b). If N_d is the donor atom concentration in the crystal, provided that $N_d \gg n_i$, then at room temperature, the electron concentration in the CB will nearly be equal to N_d, that is, $n \approx N_d$. The hole concentration will be $p = n_i^2/N_d$, which is less than the intrinsic concentration. The reason is that a small fraction of the large number of electrons in the CB recombine with holes in the VB so as to maintain $np = n_i^2$. The mass action law must be maintained in equilibrium.

The conductivity σ of a semiconductor depends on both electrons and holes as both contribute to charge transport. If μ_e and μ_h are the drift mobilities of the electrons and holes, respectively, then

Semiconductor conductivity

$$\sigma = en\mu_e + ep\mu_h \tag{3.3.1}$$

which for an *n*-type semiconductor becomes,

n-type conductivity

$$\sigma = eN_d\mu_e + e\left(\frac{n_i^2}{N_d}\right)\mu_h \approx eN_d\mu_e \tag{3.3.2}$$

We should, by similar arguments to the above, anticipate that doping a Si crystal with a trivalent atom (valency of 3) such as B (boron) will result in a *p*-type Si that has an excess of holes in the crystal. Consider doping Si with small amounts of B as shown in Figure 3.8 (a). Because B has only three valence electrons, when it shares them with four neighboring Si atoms one of the bonds has a missing electron which is of course a "hole." A nearby electron can tunnel into this hole and displace the hole further away from the B atom. As the hole moves away it gets attracted by the negative charge left behind on the B atom. The binding energy of this hole to the B^- ion (a B atom that has accepted an electron) can be calculated using the hydrogenic atom analogy just like in the *n*-type Si case. This binding energy also turns out to be very small, ~0.05 eV, so that at room temperature the thermal vibrations of the lattice can free the hole away from the B^- site. A free hole, we recall, exists in the VB. The escape of the hole from the B^- site involves the B atom accepting an electron from a neighboring Si-Si bond (from the VB) which effectively results in the hole being displaced away, and its eventual escape to freedom in the VB. The B atom introduced into the Si crystal therefore acts as an electron **acceptor** impurity. The electron accepted by the B atom comes from a nearby bond. On the energy band diagram, an electron leaves the VB and gets accepted by a B atom which becomes negatively charged. This process leaves a hole in the VB which is free to wander away as illustrated in Figure 3.8 (b).

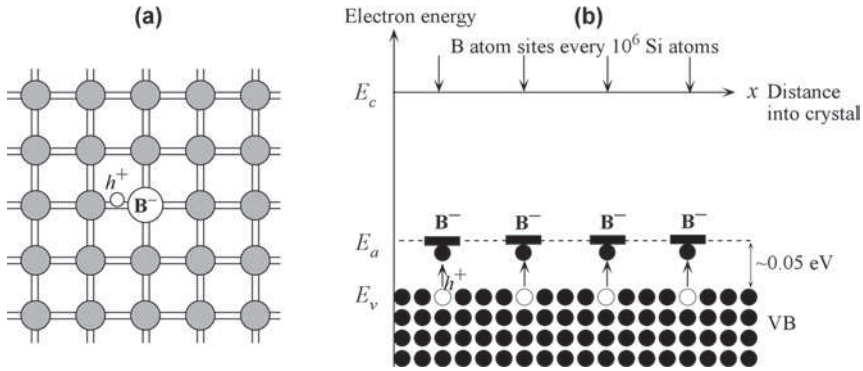


FIGURE 3.8 (a) Boron-doped Si crystal. B has only three valence electrons. When it substitutes for a Si atom one of its bonds has an electron missing and therefore a hole. (b) Energy band diagram for a *p*-type Si doped with 1 ppm B. There are acceptor energy levels just above E_v around B^- sites. These acceptor levels accept electrons from the VB and therefore create holes in the VB.

It is apparent that doping a silicon crystal with a trivalent impurity results in a *p*-type material. We have many more holes than electrons for electrical conduction since the negatively charged B atoms are immobile and hence cannot contribute to the conductivity. If the concentration of acceptor impurities N_a in the crystal is much greater than the intrinsic concentration n_i, then at room temperature all the acceptors would have been ionized and thus $p \approx N_a$. The electron concentration is then determined by the mass action law, $n = n_i^2/N_a$, which is much smaller than p, and consequently the conductivity is simply given by $\sigma = eN_a\mu_h$, similar to Eq. (3.3.2).

Figures 3.9 (a)–(c) show the energy band diagrams of an intrinsic, an *n*-type, and a *p*-type semiconductor. The energy distance of E_F from E_c and E_v determines the electron and hole concentrations by virtue of Eqs. (3.2.3) and (3.2.4). Note the locations of the Fermi level in each case: E_{Fi} for intrinsic; E_{Fn} for *n*-type; and E_{Fp} for *p*-type.

The position of the Fermi level in the energy band diagram plays an important role in understanding device principles. It is clear from Eqs. (3.2.3) and (3.2.4) that its location in the energy band diagram acts as a lever in finding the free electron and hole concentrations in the CB and VB, respectively, as highlighted in Figures 3.9 (a)–(c). It represents the electron

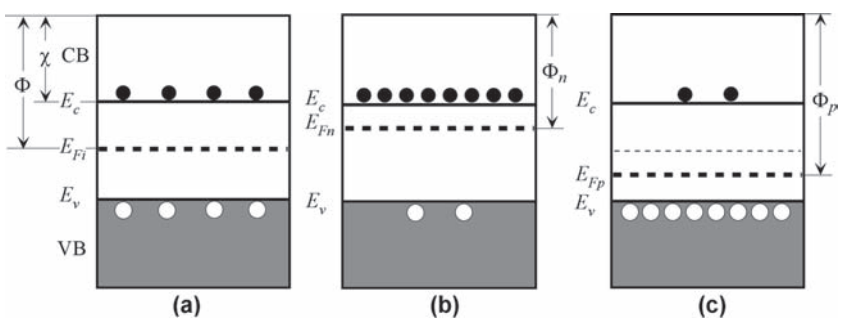


FIGURE 3.9 Energy band diagrams for (a) intrinsic, (b) *n*-type, and (c) *p*-type semiconductors. In all cases, $np = n_i^2$. Note that donor and acceptor energy levels are not shown. CB = conduction band, VB = valence band, E_c = CB edge, E_v = VB edge, E_{Fi} = Fermi level in intrinsic semiconductor, E_{Fn} = Fermi level in *n*-type semiconductor, E_{Fp} = Fermi level in *p*-type semiconductor. χ is the electron affinity. Φ, Φ_n, and Φ_p are the work functions for the intrinsic, *n*-type, and *p*-type semiconductors, respectively.

TABLE 3.1 Selected typical properties of various semiconductors at 300 K

	a (nm)	E_g (eV)	χ (eV)	N_c(cm^{-3})	N_v(cm^{-3})	n_i(cm^{-3})	ε_r	μ_e (cm^{2}V^{-1}s^{-1})	μ_h (cm^{2}V^{-1}s^{-1})
Ge (DI)	0.5650	0.66 (I)	4.13	1.04×10^{19}	6.0×10^{19}	2.3×10^{13}	16.0	3900	1900
Si (DI)	0.5431	1.11 (I)	4.05	2.8×10^{19}	1.2×10^{19}	1.0×10^{10}	11.8	1450	490
InP (ZB)	0.5868	1.35 (D)	4.50	5.2×10^{17}	1.1×10^{19}	3.0×10^{7}	12.6	4600	150
GaAs (ZB)	0.5653	1.42 (D)	4.07	4.7×10^{17}	7.0×10^{18}	2.1×10^{6}	13.0	8500	400
AlAs (ZB)	0.5661	2.17 (I)	3.50	1.5×10^{19}	1.7×10^{19}	10	10.1	200	100

Notes: Data combined from a number of sources. I and D represent indirect and direct bandgap. DI, diamond crystal; ZB, zinc blend; a, lattice constant. (Note that there are variations in the values of certain properties among books, for example, n_i for Si, ε_r for GaAs, *etc.* Most commonly used or recent values have been selected.)

and hole concentrations in a semiconductor in the energy band diagram. The work function Φ of a semiconductor is defined as the energy distance from the Fermi level to the vacuum level as shown in Figures 3.9 (a)–(c) as Φ, Φ_n, and Φ_p for the intrinsic, n-type, and p-type semiconductor, respectively. Although there are no states at E_F, the work function nonetheless represents the *average energy* required to remove an electron from a semiconductor.[3]

The following definitions and notation are generally used to describe extrinsic semiconductors. Electrons in an n-type semiconductor ($n > p$) are **majority carriers** whereas holes are **minority carriers**. The electron, that is, majority carrier, concentration in this n-type semiconductor in equilibrium is n_{no}, where the subscripts refer to the n-type semiconductor and equilibrium (which excludes photoexcitation). The hole, that is, minority carrier, concentration is denoted p_{no}. In this notation, $n_{no} = N_d$ and the mass action law is $n_{no}p_{no} = n_i^2$. Similarly, hole or majority carrier concentration in a p-type semiconductor ($p > n$) is p_{po} and electron, minority carrier, concentration is n_{po}. Further $p_{po} = N_a$ and $n_{po}p_{po} = n_i^2$.

Table 3.1 summarizes some of the properties of useful semiconductors, including Si and GaAs, at 300 K.

B. Compensation Doping

Compensation doping describes the doping of a semiconductor with both donors and acceptors to control the properties. For example, a p-type semiconductor doped with N_a acceptors can be converted to an n-type semiconductor by simply adding donors until the concentration N_d exceeds N_a. The effect of donors compensates for the effect of acceptors and vice versa. The electron concentration is then given by $N_d - N_a$ provided the latter is larger than n_i. When both acceptors and donors are present, what essentially happens is that electrons from donors recombine with the holes from the acceptors so that the mass action law $np = n_i^2$ is obeyed. Remember that we cannot simultaneously increase the electron and hole concentrations because that leads to an increase in the recombination rate which returns the electron and hole concentrations to values that satisfy $np = n_i^2$. Suppose that we have more donors than acceptors. When an acceptor atom accepts a valence band electron, a hole is created in the VB. This hole then recombines with an electron from the CB; the electron would have been supplied by a donor. If we take the initial electron concentration as $n = N_d$ then the recombination between the electrons from the donors and N_a holes generated by N_a acceptors

[3]The following intuitive argument may help. The easiest electron to remove from an intrinsic semiconductor needs an energy χ. To maintain equilibrium, we must also remove an electron from the VB, which needs an energy $E_g + \chi$ so that the average energy is $\frac{1}{2}E_g + \chi$, which is the work function of the intrinsic semiconductor.

results in the electron concentration reduced by N_a to $n = N_d - N_a$. The hole concentration is then $p = n_i^2/(N_d - N_a)$. By a similar argument, if we have more acceptors than donors, the hole concentration becomes $p = N_a - N_d$, with electrons from N_d donors recombining with holes from N_a acceptors. The electron concentration is then $n = n_i^2/(N_a - N_d)$.

C. Nondegenerate and Degenerate Semiconductors

In nondegenerate semiconductors, the number of states in the CB far exceeds the number of electrons so that the likelihood of two electrons trying to occupy the same state is almost nil. This means that the Pauli exclusion principle can be neglected and the electron statistics can be described by the Boltzmann statistics. N_c is a measure of the density of states in the CB. The Boltzmann expression in Eq. (3.2.3) for n is valid only when $n \ll N_c$. Those semiconductors for which $n \ll N_c$ and $p \ll N_v$ are termed **nondegenerate** semiconductors.

When the semiconductor has been excessively doped with donors, n may be so large, typically $10^{19}-10^{20}\,\text{cm}^{-3}$, that it may be comparable to N_c. In that case the Pauli exclusion principle becomes important in the electron statistics and we have to use the Fermi–Dirac statistics. Such a semiconductor exhibits properties that are more metal-like than semiconductor. Semiconductors that have $n \gtrsim N_c$ or $p \gtrsim N_v$ are called **degenerate semiconductors**. For degenerate semiconductors, the Fermi level is inside the band that is involved in electrical conduction. For example, the Fermi level in an n-type degenerate semiconductor is above E_c.

The large carrier concentration in a degenerate semiconductor is due to the heavy doping. For example, as the donor concentration in an n-type semiconductor is increased, at sufficiently high doping levels, the donor atoms become so close to each other that their orbitals overlap to form a narrow energy band, which overlaps and becomes part of the conduction band as illustrated in Figure 3.10 (a). This situation is reminiscent to the valence electrons filling overlapping energy bands in a metal. In a degenerate n-type semiconductor, the Fermi level is therefore within the CB, or above E_c just like E_F is within the band in a metal. Majority of the states between E_c and E_F are full of electrons as indicated in Figure 3.10 (a). Notice that E_c has been pushed down so that the bandgap is *narrower* in heavily doped semiconductors. The random distribution of a large number of donors in the crystal also introduces a random variation in the potential energy (PE) of the electron; the PE is no longer perfectly periodic. Put differently, there are spatial variations in the density of states in the crystal. Such random fluctuations in the PE give rise to a small tail (an extension) in the density of states $g(E)$ in the bandgap, called **band tailing**, so that E_c is no longer sharp in heavily doped crystals.

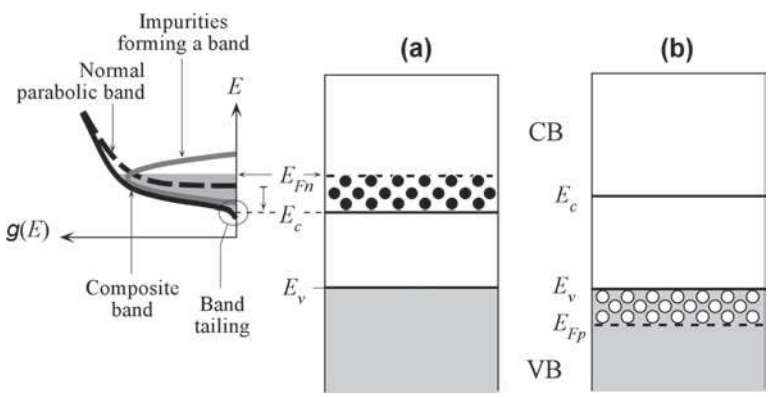


FIGURE 3.10
(a) Degenerate n-type semiconductor. Large number of donors form a band that overlaps the CB. E_c is pushed down and E_{Fn} is within the CB. (b) Degenerate p-type semiconductor.

In the case of a *p*-type degenerate semiconductor, the Fermi level lies in the VB below E_v as in Figure 3.10 (b). One cannot simply assume that $n = N_d$ or $p = N_a$ in a degenerate semiconductor because the dopant concentration is so large that they interact with each other. Not all dopants are able to become ionized and the carrier concentration eventually reaches a saturation typically around ~10^{20} cm^{-3}. Furthermore, the mass action law, $np = n_i^2$, is not valid for degenerate semiconductors.

D. Energy Band Diagrams in an Applied Field

Consider an *n*-type semiconductor that is connected to a voltage supply of V and is carrying a current as shown in Figure 3.11. From basic electrical conduction we know that there is an electric field E inside the semiconductor that drifts the conduction electrons and hence causes a current flow. Consider now the energy band diagram. The Fermi level E_F is above that for the intrinsic case (E_{Fi}), closer to E_c than E_v. The applied voltage drops uniformly along the semiconductor so that the electrons in the semiconductor now also have an imposed electrostatic potential energy, which decreases toward the positive terminal as illustrated in Figure 3.11. The whole band structure, the CB and the VB, therefore tilts. When an electron drifts from A toward B, its PE decreases because it is approaching the positive terminal.

For a semiconductor system in the dark, in equilibrium and with no applied voltage or no EMF generated, E_F must be uniform across the system since $\Delta E_F = eV = 0$. However, when electrical work is done on the system, for example, when a battery is connected to a semiconductor, E_F is not uniform throughout the whole system. A change ΔE_F in E_F within a material system is equivalent to electrical work per electron or eV. The Fermi level E_F therefore follows the electrostatic PE behavior. The change in E_F from one end to the other, $E_F(A) - E_F(B)$, is just eV, the energy needed in taking an electron through the semiconductor as shown in Figure 3.11. The electron concentration in the semiconductor is uniform so that $E_c - E_F$ must be constant from one end to the other. Thus the CB, VB, and E_F all bend by the same amount.

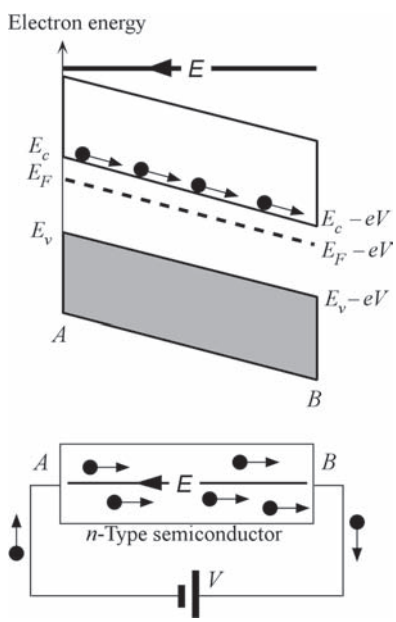


FIGURE 3.11 Energy band diagram of an *n*-type semiconductor connected to a voltage supply of V volts. The whole energy diagram tilts because the electron now has an electrostatic potential energy as well.

EXAMPLE 3.3.1 Fermi levels in semiconductors

An *n*-type Si wafer has been doped uniformly with 10^{16} phosphorus (P) atoms cm^{-3}. Calculate the position of the Fermi energy with respect to the Fermi energy E_{Fi} in intrinsic Si. The above *n*-type Si sample is further doped with 2×10^{17} boron atoms cm^{-3}. Calculate the position of the Fermi energy with respect to the Fermi energy E_{Fi} in intrinsic Si at room temperature (300 K), and hence with respect to the Fermi energy in the *n*-type case above.

Solution

P (Group V) gives *n*-type doping with $N_d = 10^{16}\,cm^{-3}$, and since $N_d \gg n_i\,(=10^{10}\,cm^{-3}$ from Table 3.1), we have $n = N_d = 10^{16}\,cm^{-3}$. For intrinsic Si

$$n_i = N_c \exp\left[-(E_c - E_{Fi})/k_B T\right]$$

whereas for doped Si,

$$n = N_c \exp\left[-(E_c - E_{Fn})/k_B T\right] = N_d$$

where E_{Fi} and E_{Fn} are the Fermi energies in the intrinsic and *n*-type Si. Dividing the two expressions,

$$N_d/n_i = \exp\left[(E_{Fn} - E_{Fi})/k_B T\right]$$

so that

$$E_{Fn} - E_{Fi} = k_B T \ln(N_d/n_i) = (0.0259\,eV)\ln(10^{16}/10^{10}) = 0.358\,eV$$

When the wafer is further doped with boron, the acceptor concentration, $N_a = 2 \times 10^{17}\,cm^{-3} > N_d = 10^{16}\,cm^{-3}$. The semiconductor is compensation doped and compensation converts the semiconductor to a *p*-type Si. $p = N_a - N_d = 2 \times 10^{17} - 10^{16} = 1.9 \times 10^{17}\,cm^{-3}$.

For intrinsic Si

$$p = n_i = N_v \exp\left[-(E_{Fi} - E_v)/k_B T\right],$$

whereas for doped Si

$$p = N_v \exp\left[-(E_{Fp} - E_v)/k_B T\right] = N_a - N_d$$

where E_{Fi} and E_{Fp} are the Fermi energies in the intrinsic and *p*-type Si, respectively. Dividing the two expressions

$$p/n_i = \exp\left[-(E_{Fp} - E_{Fi})/k_B T\right]$$

so that

$$E_{Fp} - E_{Fi} = -k_B T \ln(p/n_i) = -(0.0259\,eV)\ln(1.9 \times 10^{17}/1.0 \times 10^{10}) = -0.434\,eV$$

The negative value indicates that E_{Fp} is below E_{Fi}.

EXAMPLE 3.3.2 Conductivity of *n*-Si

Consider a pure intrinsic Si crystal. What would be its intrinsic conductivity at 300 K? What are the electron and hole concentrations in an *n*-type Si crystal that has been doped with $10^{16}\,cm^{-3}$ phosphorus (P) donors. What is the conductivity if the drift mobility of electrons is about 1200 $cm^2\,V^{-1}\,s^{-1}$ at this concentration of dopants?

Solution

The intrinsic concentration $n_i = 1 \times 10^{10}\,\text{cm}^{-3}$, so that the intrinsic conductivity is

$$\sigma = en_i(\mu_e + \mu_h) = (1.6 \times 10^{10}\,\text{C})(1 \times 10^{10}\,\text{cm}^{-3})(1450 + 490\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1})$$

$$= 3.1 \times 10^{-6}\,\Omega^{-1}\,\text{cm}^{-1} \quad \text{or} \quad 3.1 \times 10^{-4}\,\Omega^{-1}\,\text{m}^{-1}$$

Consider *n*-type Si. $N_d = 10^{16}\,\text{cm}^{-3} > n_i(=10^{10}\,\text{cm}^{-3})$, the electron concentration $n = N_d = 10^{16}\,\text{cm}^{-3}$ and $p = n_i^2/N_d = (10^{10}\,\text{cm}^{-3})^2/(10^{16}\,\text{cm}^{-3}) = 10^4\,\text{cm}^{-3}$; and negligible compared to *n*.
The conductivity is

$$\sigma = eN_d\mu_e = (1.6 \times 10^{-19}\,\text{C})(1 \times 10^{16}\,\text{cm}^{-3})(1200\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}) = 1.92\,\Omega^{-1}\,\text{cm}^{-1}$$

3.4 DIRECT AND INDIRECT BANDGAP SEMICONDUCTORS: *E–k* DIAGRAMS

We know from quantum mechanics that when the electron is within an infinite potential energy well of spatial width *L*, its energy is quantized and given by

$$E_n = \frac{(\hbar k_n)^2}{2m_e}$$

where m_e is the mass of the electron and the electron's wave vector k_n is essentially a quantum number determined by

$$k_n = \frac{n\pi}{L}$$

where $n = 1, 2, 3, \ldots$. The energy increases parabolically with the wave vector k_n. We also know that the electron momentum is given by $\hbar k_n$. This description can be used to represent the behavior of electrons in a metal within which their average potential energy can roughly be taken to be zero. In other words, we take $V(x) = 0$ within the metal crystal and $V(x)$ to be large, for example, $V(x) = V_o$ (several electron volts) outside, so that the electron is contained within the metal. This is the **nearly free electron model** of a metal which has been quite successful in interpreting many of the metallic properties. Indeed, we typically calculate the density of states $g(E)$ based on the three-dimensional potential well problem. However, it is quite obvious that this model is too simple because it does not take into account the actual variation of the electron potential energy in the crystal.

The potential energy of the electron depends on its location within the crystal and is periodic due to the regular arrangement of the atoms. How does a periodic potential energy affect the relationship between *E* and *k*? It will no longer be simply $E_n = (\hbar k_n)^2/2m_e$.

To find the energy of the electron in a crystal, we need to solve the Schrödinger equation for a periodic potential energy function in three dimensions. We first consider the hypothetical one-dimensional crystal shown in Figure 3.12. The electron potential energy functions for each atom add to give an overall potential energy function $V(x)$, which is clearly periodic in *x* with the periodicity of the crystal, *a*. Thus, $V(x) = V(x + a) = V(x + 2a) = \ldots$ and so on. Our task is therefore to solve the Schrödinger equation for the electron wavefunction $\psi(x)$,

$$\frac{d^2\psi}{dx^2} + \frac{2m_e}{\hbar^2}\left[E - V(x)\right]\psi = 0 \qquad \textbf{(3.4.1)}$$

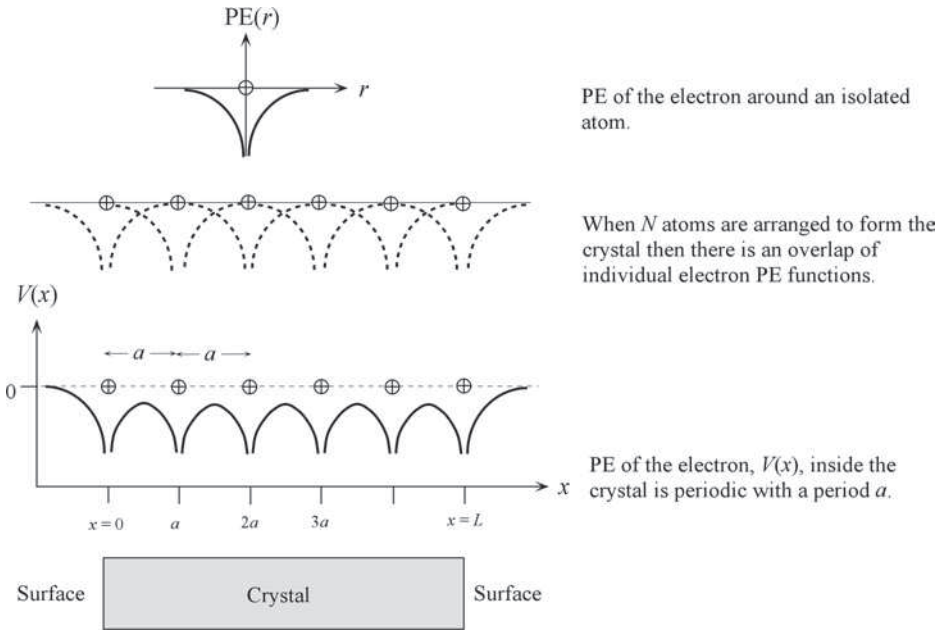


FIGURE 3.12 The electron potential energy (PE), $V(x)$, inside the crystal is periodic with the same periodicity as that of the crystal, a. Far away outside the crystal, by choice, $V = 0$ (the electron is free and PE $= 0$).

subject to the condition that the potential energy, $V(x)$, is periodic in a, that is,

$$V(x) = V(x + ma); \quad m = 1, 2, 3, \ldots \tag{3.4.2}$$

Periodic PE in the crystal

The solution of Eq. (3.4.1) will give the electron wavefunction in the crystal and hence the electron energy. Since $V(x)$ is periodic, we should expect, by intuition at least, the solution $\psi(x)$ to be periodic. It turns out that the solutions to Eq. (3.4.1), which are called **Bloch wavefunctions**, are of the form

$$\psi_k(x) = U_k(x)\exp(jkx) \tag{3.4.3}$$

Bloch wave in the crystal

where $U_k(x)$ is a periodic function which depends on $V(x)$ and has the same periodicity a as $V(x)$. The term $\exp(jkx)$, of course, represents a traveling wave whose wave vector is k. We should remember that we have to multiply this by $\exp(-jEt/\hbar)$, where E is the energy, to get the overall wavefunction $\Psi(x, t)$. Thus, the electron wavefunction in the crystal is a traveling wave which is modulated by $U_k(x)$. Further, both $\exp(jkx)$ and $\exp(-jkx)$ are possible and represent left and right traveling waves.

There are many such Bloch wavefunction solutions to the one-dimensional crystal, each identified with a particular k value, say k_n, which acts as a kind of quantum number. Each $\psi_k(x)$ solution corresponds to a particular k_n and represents **a state** with an energy E_k. The dependence of the energy E_k on the wave vector k is illustrated in what is called an ***E–k* diagram**. Figure 3.13 shows a typical *E–k* diagram for the hypothetical one-dimensional solid for k values in the range $-\pi/a$ to $+\pi/a$. Just as $\hbar k$ is the momentum of a free electron, $\hbar k$ for the Bloch electron is the momentum involved in its interaction with external fields, for example, those involved in the photon absorption processes. Indeed the rate of change of $\hbar k$ is the externally applied force F_{ext} on

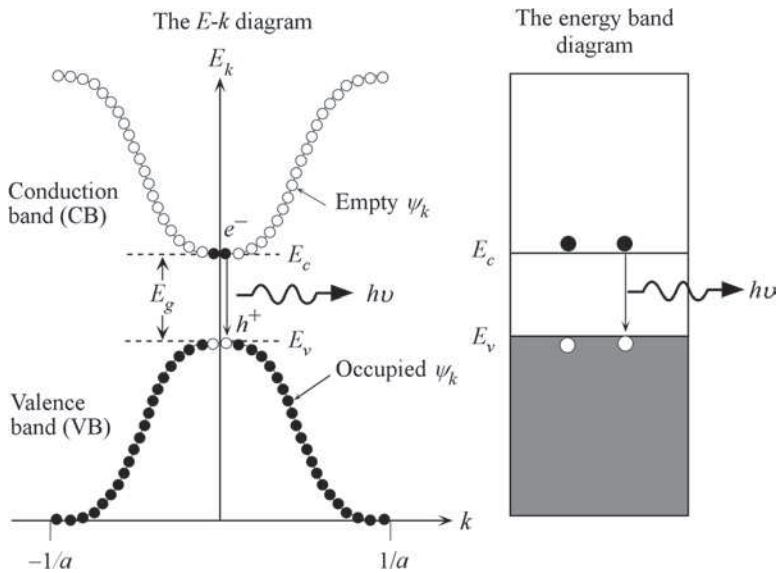


FIGURE 3.13 The *E–k* diagram of a direct bandgap semiconductor such as GaAs. The *E–k* curve consists of many discrete points with each point corresponding to a possible state, wavefunction $\psi_k(x)$, that is allowed to exist in the crystal. The points are so close that we normally draw the *E–k* relationship as a continuous curve. In the energy range E_v to E_c there are no points, that is, no $\psi_k(x)$ solutions.

the electron such as that due to an electric field E (*i.e.*, $F_{\text{ext}} = eE$). Thus, for the electron within the crystal, $d(\hbar k)/dt = F_{\text{ext}}$ and consequently we call $\hbar k$ the **crystal momentum** of the electron.[4]

Inasmuch as the momentum of the electron in the *x*-direction in the crystal is given by $\hbar k$, the *E–k* diagram is an energy vs. crystal momentum plot. The states $\psi_k(x)$ in the lower *E–k* curve constitute the wavefunctions for the valence electrons and thus correspond to the states in the valence band. Those in the upper *E–k* curve, on the other hand, correspond to the states in the conduction band since they have higher energies. All the valence electrons at 0 K therefore fill the states (particular k_n values) in the lower *E–k* diagram.

It should be emphasized that an *E–k* curve in the diagram consists of many discrete points, each point corresponding to a possible state, wavefunction $\psi_k(x)$, that is allowed to exist in the crystal. The points are so close that we draw the *E–k* relationship as a continuous curve. It is clear from the *E–k* diagram that there is a range of energies, from E_v to E_c, for which there are no solutions to the Schrödinger equation and hence there are no $\psi_k(x)$ with energies in E_v to E_c. Furthermore, we also note that the *E–k* behavior is not a simple parabolic relationship except near the bottom of the CB and the top of the VB.

Above absolute zero of temperature, due to thermal excitation, however, some of the electrons from the top of the valence band will be excited to the bottom of the conduction band. According to the *E–k* diagram in Figure 3.13, when an electron and hole recombine, the electron simply drops from the bottom of the CB to the top of the VB without any change in its *k* value so that this transition is quite acceptable in terms of momentum conservation. We should recall that the momentum of the emitted photon is negligible compared with the momentum of the electron. The *E–k* diagram in Figure 3.13 is therefore for a **direct bandgap semiconductor**. The minimum of the CB is directly above the maximum of the VB. As shown in Figure 3.13, the electron transition from the bottom of the CB to the top of the VB emits a photon.

---

[4]The actual momentum of the electron, however, is not $\hbar k$ because $d(\hbar k)/dt \neq F_{\text{external}} + F_{\text{internal}}$. The true momentum p_e satisfies $dp_e/dt = F_{\text{external}} + F_{\text{internal}}$ (all forces on the electron). However, as we are interested in interactions with external forces such as an applied field, we treat $\hbar k$ as if it were the momentum of the electron in the crystal and use the name *crystal momentum.*

The simple *E–k* diagram sketched in Figure 3.13 is for the hypothetical one-dimensional crystal in which each atom simply bonds with two neighbors. In real crystals, we have a three-dimensional arrangement of atoms with $V(x, y, z)$ showing periodicity in more than one direction. The *E–k* curves are then not as simple as that in Figure 3.13 and often show unusual features. The *E–k* diagram for GaAs, which is shown in Figure 3.14 (a), as it turns out, has general features that are quite similar to that sketched in Figure 3.13. The lowest CB minimum is right above the top of the VB at the same *k*-value. GaAs is therefore a direct bandgap semiconductor in which electron–hole pairs can recombine directly and emit a photon. The majority of light-emitting devices use direct bandgap semiconductors to make use of direct recombination.

In the case of Si, the diamond crystal structure leads to an *E–k* diagram which has the essential features illustrated in Figure 3.14 (b). We notice that the minimum of the CB is *not* directly above the maximum of the VB, but it is displaced on the *k*-axis. Such crystals are called **indirect bandgap semiconductors**. An electron at the bottom of the CB cannot therefore recombine directly with a hole at the top of the VB, because for the electron to fall down to the top of the VB its momentum must change from k_{CB} to k_{VB}, which is not allowed by the law of conservation of momentum. Thus, direct electron–hole recombination does not normally take place in Si and Ge. The recombination process in these elemental semiconductors occurs via a **recombination center** at an energy level E_r within the bandgap as illustrated in Figure 3.14 (c). These recombination centers may be crystal defects or impurities. The electron is first captured by the defect at E_r. The change in the energy and momentum of the electron by this capture process is transferred to lattice vibrations, that is, to **phonons**. As much as an electromagnetic radiation is quantized in terms of photons, lattice vibrations in the crystal are quantized in terms of phonons. Lattice vibrations travel in the crystal just like a wave and these waves are called phonons. The captured electron at E_r can readily fall down into an empty state at the top of the VB and thereby recombine with a hole as in Figure 3.14 (c). Typically, the electron transition from E_r to E_v involves the emission of further lattice vibrations.

In some indirect bandgap semiconductors such as GaP, however, the recombination of the electron with a hole at certain recombination centers results in photon emission. The *E–k* diagram is similar to that shown in Figure 3.14 (c) except that the recombination centers at E_r are generated by the purposeful addition of nitrogen (N) impurities to GaP, written as GaP:N. The electron transition from E_r to E_v involves photon emission in the green.

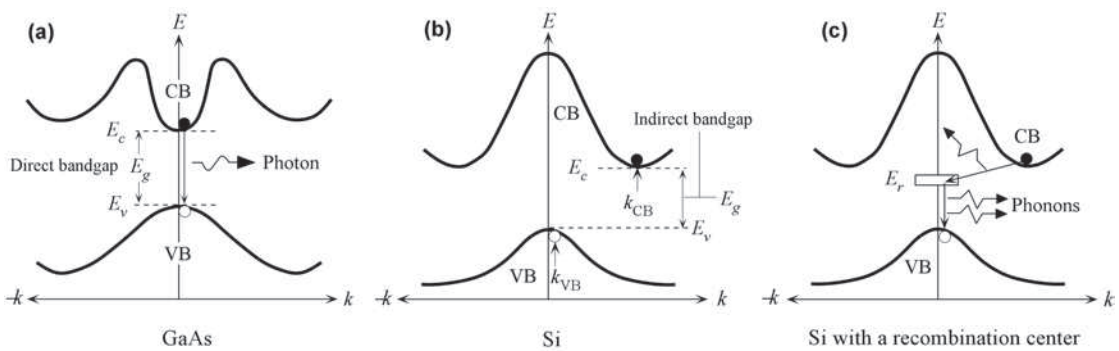


FIGURE 3.14 (a) In GaAs the minimum of the CB is directly above the maximum of the VB. GaAs is therefore a direct bandgap semiconductor. (b) In Si, the minimum of the CB is displaced from the maximum of the VB and Si is an indirect bandgap semiconductor. (c) Recombination of an electron and a hole in Si involves a recombination center.

3.5 *pn* JUNCTION PRINCIPLES

A. Open Circuit

Consider what happens when one side of a sample of Si is doped *n*-type and the other *p*-type as shown in Figure 3.15 (a). Assume that there is an *abrupt* discontinuity between the *p* and *n* regions which we call the **metallurgical junction**, *M*, as indicated in Figure 3.15 (a). There are fixed (immobile) ionized donors and free electrons (in the conduction band, CB) in the *n*-region and fixed ionized acceptors and holes (in the valence band, VB) in the *p*-region.

Due to the hole concentration gradient from the $p = p_{po}$ to the *n*-side where $p = p_{no}$, holes *diffuse* toward the right and enter the *n*-region and recombine with the electrons (majority carriers)

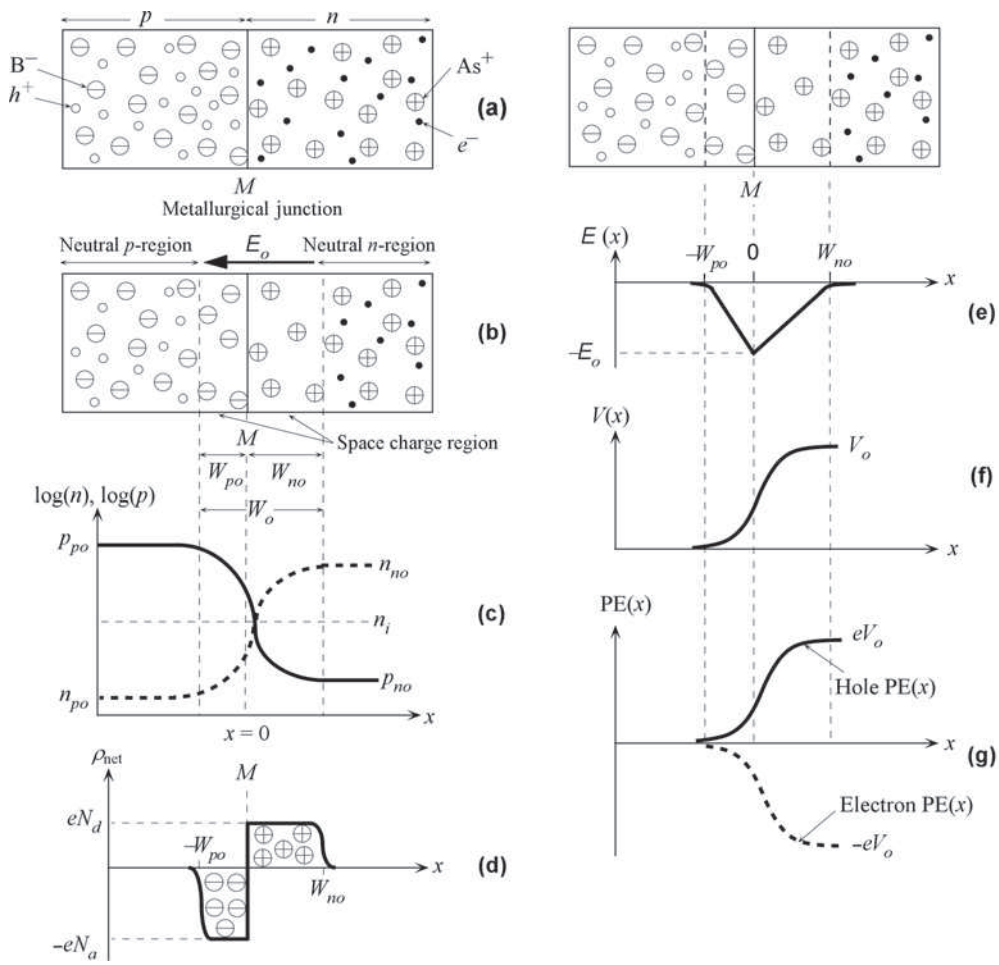


FIGURE 3.15 Properties of the *pn* junction. (a) The *p*- and *n*-sides of the *pn* junction before the contact. (b) The *pn* junction after contact, in equilibrium and in open circuit. (c) Carrier concentrations along the whole device through the *pn* junction. At all points, $n_{po}p_{po} = n_{no}p_{no} = n_i^2$. (d) Net space charge density ρ_{net} across the *pn* junction. (e) The electric field across the *pn* junction is found by integrating ρ_{net} in (d). (f) The potential $V(x)$ across the device. Contact potentials are not shown at the semiconductor–metal contacts. (g) Hole and electron potential energy (PE) across the *pn* junction. Potential energy is charge × potential = qV.

in this region. The *n*-side near the junction therefore becomes depleted of majority carriers and therefore has exposed positive donor ions (As^+) of concentration N_d. Similarly, the electron concentration gradient drives the electrons by diffusion toward the left. Electrons diffusing into the *p*-side recombine with the holes (majority carriers), which exposes negative acceptor ions (B^-) of concentration N_a in this region. The regions on both sides of the junction *M* consequently become *depleted* of free carriers in comparison with the bulk *p*- and *n*-regions far away from the junction. There is therefore a **space charge layer** (SCL) around *M*. Figure 3.15 (b) shows the SCL, also known as the **depletion region**, around *M*. Figure 3.15 (c) illustrates the hole and electron concentration profiles in which the vertical concentration scale is logarithmic. Note that we must, under equilibrium conditions (*e.g.*, no applied bias or photoexcitation), have $pn = n_i^2$ everywhere. The depletion region extends from about $x = -W_{po}$ to $x = W_{no}$. (The subscripts indicate the *p*- or *n*-side and open circuit.). The total width of the depletion region is $W_o = W_{po} + W_{no}$.

It is clear that there is an internal electric field E_o from positive ions to negative ions, that is, in the $-x$ direction, which tries to drift the holes back into the *p*-region and electrons back into the *n*-region. This field drives the holes in the opposite direction to their diffusion. As shown in Figure 3.15 (b), E_o imposes a drift force on holes in the $-x$ direction whereas the hole diffusion flux is in the $+x$ direction. A similar situation also applies for electrons with the electric field attempting to drift the electrons against diffusion from *n*- to the *p*-region. It is apparent that as more and more holes diffuse toward the right, and electrons toward the left, the internal field around *M* will increase until eventually an *equilibrium* is reached when the rate of holes diffusing toward the right is just balanced by the rates of holes drifting back to the left, driven by the field E_o. The electron diffusion and drift fluxes will also be balanced in equilibrium. For uniformly doped *p*- and *n*-regions, the net space charge density $\rho_{\text{net}}(x)$ across the semiconductor will be as shown in Figure 3.15 (d). The net space charge density ρ_{net} is negative and equal to $-eN_a$ in the SCL from $x = -W_{po}$ to $x = 0$ (*M* is at $x = 0$) and then positive and equal to $+eN_d$ from $x = 0$ to W_{no}. The total charge on the left-hand side must equal to that on the right-hand side for overall charge neutrality, so that

$$N_a W_{po} = N_d W_{no} \tag{3.5.1}$$

Depletion region widths

Figure 3.15 arbitrarily assumes that the acceptor concentration is greater than the donor concentration, $N_a > N_d$. From Eq. (3.5.1) this implies that $W_{no} > W_{po}$, that is, the depletion region penetrates the *n*-side (lightly doped side) more than the *p*-side (heavily doped side). Indeed, if $N_a \gg N_d$, then the depletion region is almost entirely on the *n*-side. We generally indicate heavily doped regions with the superscript plus sign as p^+. The electric field $E(x)$ and the net space charge density $\rho_{\text{net}}(x)$ at a point are related in electrostatics[5] by $dE/dx = \rho_{\text{net}}(x)/\varepsilon$, where $\varepsilon = \varepsilon_o \varepsilon_r$, is the permittivity of the medium, and ε_o and ε_r are the absolute permittivity and relative permittivity of the semiconductor material, respectively. We can thus integrate $\rho_{\text{net}}(x)$ across the device and thus determine the electric field. The variation of the electric field across the *pn* junction is shown in Figure 3.15 (e). The negative field means that it is in the $-x$ direction. Note that $E(x)$ reaches a maximum magnitude E_o at *M*.

The potential $V(x)$ at any point x can be found by integrating the electric field since by definition $E = -dV/dx$. Taking the potential on the *p*-side far away from *M* as zero (we have no applied voltage), which is an arbitrary reference level, $V(x)$ increases in the depletion region

[5]This is called *Gauss's law in point* form and comes from Gauss's law in electrostatics. The integration of the electric field *E* over a closed surface *S* is related to the total charge Q_{enclosed} enclosed, $\int E dS = Q_{\text{enclosed}}/\varepsilon$, where ε is the permittivity of the medium.

toward the *n*-side as indicated in Figure 3.15 (f). Notice that on the *n*-side the potential reaches V_o, which is called the **built-in potential**.

In an abrupt *pn* junction $\rho_{\text{net}}(x)$ can simply and approximately be described by step functions as displayed in Figure 3.15 (d). Using the step form of $\rho_{\text{net}}(x)$ in Figure 3.15 (d) and integrating it gives the electric field E_o and the built-in potential V_o,

Built-in field

$$E_o = -\frac{eN_dW_{no}}{\varepsilon} = -\frac{eN_aW_{po}}{\varepsilon} \tag{3.5.2}$$

and

Built-in potential

$$V_o = -\frac{1}{2}E_oW_o = -\frac{eN_aN_dW_o^2}{2\varepsilon(N_a + N_d)} \tag{3.5.3}$$

where $\varepsilon = \varepsilon_o\varepsilon_r$ is the permittivity of the semiconductor medium and $W_o = W_{no} + W_{po}$ is the total width of the depletion region under open circuit. If we know W_o, then W_{no} or W_{po} follows readily from Eq. (3.5.1). Equation (3.5.3) is a relationship between the built-in voltage V_o and the depletion region width W_o. If we know V_o we can calculate W_o.

The simplest way to relate V_o to the doping parameters is through Boltzmann statistics. For the system consisting of *p*- and *n*-type semiconductors together (forming one system), in equilibrium, the Boltzmann statistics[6] demands that the concentrations n_1 and n_2 of carriers at potential energies E_1 and E_2 be related by

Boltzmann statistics

$$n_2/n_1 = \exp\left[-(E_1 - E_2)/k_BT\right] \tag{3.5.4}$$

in which E is the potential energy—that is, qV, where q is charge and V voltage. Considering electrons, $q = -e$, we see from Figure 3.15 (g) that $E = 0$ on the *p*-side far away from M where $n = n_{po}$; and $E = -eV_o$ on the *n*-side away from M where $n = n_{no}$. Thus,

$$n_{po}/n_{no} = \exp(-eV_o/k_BT) \tag{3.5.5a}$$

Equation (3.5.5a) shows that V_o depends on n_{no} and n_{po} and hence on N_d and N_a. The corresponding equation for hole concentrations is

$$p_{no}/p_{po} = \exp(-eV_o/k_BT) \tag{3.5.5b}$$

Thus, rearranging Eqs. (3.5.5a) and (3.5.5b) we obtain

$$V_o = (k_BT/e)\ln(n_{no}/n_{po}) \quad \text{and} \quad V_o = (k_BT/e)\ln(p_{po}/p_{no})$$

We can now write p_{po} and p_{no} in terms of the dopant concentrations inasmuch as $p_{po} = N_a$, $p_{no} = n_i^2/n_{no} = n_i^2/N_d$, so that V_o becomes

Built-in potential

$$V_o = \frac{k_BT}{e}\ln\left(\frac{N_aN_d}{n_i^2}\right) \tag{3.5.6}$$

---

[6] We use Boltzmann statistics, *i.e.*, $n(E) \propto \exp(-E/k_BT)$, because the concentration of electrons in the conduction band whether on the *n*- or *p*-side is never so large that the Pauli exclusion principle becomes important. As long as the carrier concentration in the conduction band is much smaller than N_c, we can use Boltzmann statistics.

Clearly V_o has been conveniently related to the dopant and parent material properties via N_a, N_d, and n_i^2, which is given by $(N_c N_v)\exp(-E_g/k_B T)$. The built-in voltage ($V_o$) is the potential across a *pn* junction, going from *p*- to *n*-type semiconductor, in an open circuit. V_o is not the actual voltage across the diode. The voltage across the diode is made up of V_o and the contact potentials at the metal-to-semiconductor junctions at the electrodes. If we add V_o and the contact potentials at the electrode ends, we will find zero. Once we know the built-in potential V_o from Eq. (3.5.6), we can then calculate the width W_o of the depletion region from Eq. (3.5.3). The term $k_B T/e$ in Eq. (3.5.6) is called the **thermal voltage** V_{th}, and is about 26 mV at 300 K.

B. Forward Bias and the Shockley Diode Equation

Consider what happens when a battery with a voltage V is connected across a *pn* junction so that the positive terminal of the battery is attached to the *p*-side and the negative terminal to the *n*-side (forward bias). The negative polarity of the supply will reduce the potential barrier V_o by V, as shown in Figure 3.16 (a). The reason is that the bulk regions outside the SCL have high conductivities, due to the plenty of majority carriers in the bulk, in comparison with the depletion region in which there are mainly immobile ions. Thus, the applied voltage drops mostly across the depletion width W. The applied bias V now directly opposes V_o. The potential barrier against diffusion therefore becomes reduced to $(V_o - V)$ as illustrated in Figure 3.16 (b). This has drastic consequences because the probability that a hole in the *p*-side will surmount this potential barrier and diffuse to the *n*-side now becomes proportional to $\exp\left[-e(V_o - V)/k_B T\right]$. In other words, the applied voltage effectively reduces the built-in potential and hence the built-in field which acts against diffusion. Consequently, many holes can now diffuse across the depletion region and enter the *n*-side. This results in the **injection of excess minority carriers**, that is, holes into the *n*-region. Similarly, excess electrons can now diffuse toward the *p*-side and enter this region and thereby become injected minority carriers.

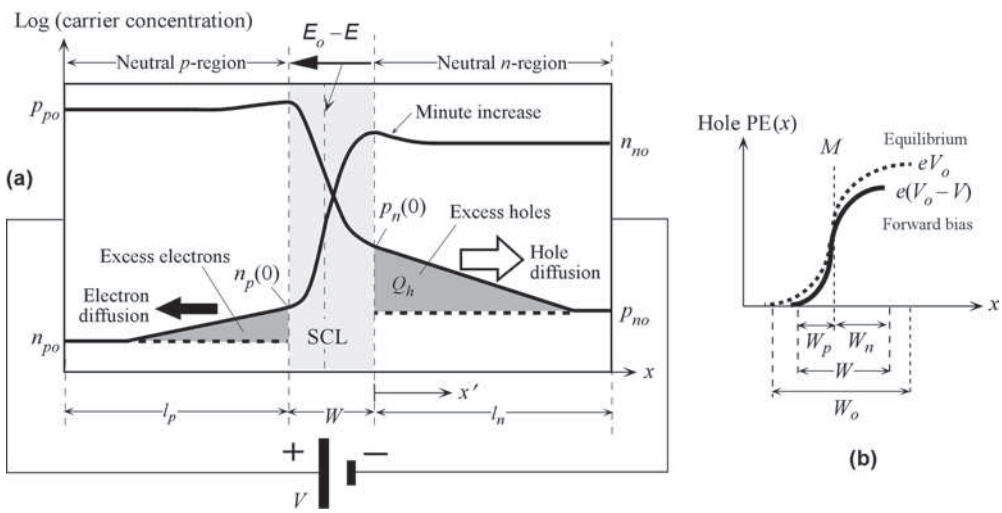


FIGURE 3.16 Forward-biased *pn* junction and the injection of minority carriers (a) Carrier concentration profiles across the device under forward bias. (b) The hole potential energy with and without an applied bias. W is the width of the SCL with forward bias.

The hole concentration $p_n(0)$ just outside the depletion region at $x' = 0$ (x' is measured from W_n) in the n-side is very large due to the injection of minority carriers as shown in Figure 3.16 (a). The hole concentration at the end of the neutral n-side at $x' = l_n$ at the negative terminal is always the equilibrium concentration p_{no}, because any additional holes at $x' = l_n$ would be neutralized immediately by electrons flowing from the negative terminal of the battery. It is obvious that there is a very large hole concentration gradient in the n-side as a result of minority carrier injection under forward bias, as shown in Figure 3.16 (a). The hole concentration gradient causes hole diffusion flux toward the negative terminal, and hence gives rise to an electric current. There is a similar electron diffusion flux and hence an electric current in the p-side. These two minority carrier diffusion fluxes constitute the forward bias pn junction current.

When holes are injected into the neutral n-side, they draw some electrons from the bulk of n-side (and hence from the battery) so that there is a small increase in the electron concentration. This small increase in the majority carriers is necessary to balance the hole charges and maintain neutrality in the n-side.

We now derive the pn junction equation. Consider the hole concentration profile in the n-side. The hole concentration, $p_n(0) = p_n(x' = 0)$, just outside the depletion region is due to the excess of holes diffusing through the SCL as a result of the reduction in the built-in potential barrier. This concentration, $p_n(0)$, is determined by the probability of surmounting the new potential energy barrier $e(V_o - V)$,

Boltzmann statistics and applied voltage

$$p_n(0) = p_{po} \exp\left[-\frac{e(V_o - V)}{k_B T}\right] \tag{3.5.7}$$

This follows directly from the Boltzmann equation, by virtue of the hole potential energy changing by $e(V_o - V)$ from $x = -W_p$ to $x = W_n$, as indicated in Figure 3.16 (b), and at the same time the hole concentration falling from p_{po} to $p_n(0)$. By dividing Eq. (3.5.7) by Eq. (3.5.5b), we get the effect of the applied voltage out directly, which shows how the voltage V determines the amount of *excess* holes diffusing and arriving at the n-region,

Law of the junction

$$p_n(0) = p_{no} \exp\left(\frac{eV}{k_B T}\right) \tag{3.5.8}$$

which is called the **law of the junction**. Equation (3.5.8) describes the effect of the applied voltage V on the injected minority carrier concentration just outside the depletion region, $p_n(0)$. Obviously, with no applied voltage, $V = 0$ and $p_n(0) = p_{no}$ which is exactly what we expect, Eq. (3.5.5b).

Injected holes diffuse in the n-region and eventually recombine with electrons in this region; there are many electrons in the n-side. Those electrons lost by recombination are readily replenished by the negative terminal of the battery connected to this side. The current due to holes diffusing in the n-region can be sustained because more holes can be supplied by the p-region, which itself can be replenished by the positive terminal of the battery.

Electrons are similarly injected from the n-side to the p-side. The electron concentration $n_p(0)$ just outside the depletion region at $x = -W_p$ is given by the equivalent of Eq. (3.5.8) for electrons, that is

Law of the junction

$$n_p(0) = n_{po} \exp\left(\frac{eV}{k_B T}\right) \tag{3.5.9}$$

In the *p*-region, the injected electrons diffuse toward the positive terminal looking to be collected. As they diffuse, they recombine with some of the many holes in this region. Those holes lost by recombination can be readily replenished by the positive terminal of the battery connected to this side. The current due to the diffusion of electrons in the *p*-side can be maintained by the supply of electrons from the *n*-side, which itself can be replenished by the negative terminal of the battery. It is apparent that an electric current can be maintained through a *pn* junction under forward bias, and that the current flow seems to be due to the *diffusion of minority carriers.* There is, however, some drift of majority carriers as well. The semiconductor *pn* junction is a minority carrier device in the sense that minority carriers play an important role.

When a hole (minority carrier) is injected into the *n*-side, it will diffuse in this region inasmuch as there is very little electric field in the neutral *n*-region to give rise to drift. Eventually the hole will recombine with an electron (majority carrier). The average time it takes for a hole (minority carrier) to recombine with an electron in the *n*-side is called the **minority carrier recombination time** or **lifetime** τ_h. The reciprocal $1/\tau_h$ is the mean probability per unit time that a hole will recombine and disappear. The ability of holes to diffuse in a semiconductor is represented by their **diffusion coefficient** D_h, which is related the their drift mobility μ_h through the Einstein relation.[7] An average distance a hole diffuses in the *n*-side before it recombines with an electron is called the **diffusion length** L_h and is given by $L_h = (D_h\tau_h)^{1/2}$. We can interpret the reciprocal $1/L_h$ in a similar way to $1/\tau_h$ as the probability per unit distance that a hole recombines with an electron in the *n*-side as it diffuses along *x*.

If the length of the *n*-region l_n is much longer than the minority carrier diffusion length L_h, most (nearly all) injected holes will eventually recombine before they reach the negative electrode. The hole concentration profile $p_n(x')$ vs. x' therefore decays from $p_n(0)$ toward the thermal equilibrium value, p_{no}, as illustrated in Figure 3.16 (a). Consider $\Delta p_n(x') = p_n(x') - p_{no}$, which is the **excess carrier concentration**, then the change $\delta\Delta p_n$ in excess concentration over a distance δx due to recombination would depend on Δp_n itself (how many excess holes we have) and also on the probability of recombination $(\delta x')(1/L_h)$ so that $\delta\Delta p_n = -\Delta p_n(\delta x'/L_h)$, which implies an exponential decay,

Long diode minority carrier profile

$$\Delta p_n(x') = \Delta p_n(0)\exp(-x'/L_h) \tag{3.5.10}$$

Equation (3.5.10) can be derived by rigorous means using the Continuity Equation[8] rather than the above intuitive arguments. The *pn* junction in which $l_n \gg L_h$ is called a **long diode**.

The hole diffusion current density $J_{D,\text{hole}}$ is the *hole diffusion flux* multiplied by the hole charge,[9]

$$J_{D,\text{hole}} = -eD_h\frac{dp_n(x')}{dx'} = -eD_h\frac{d\Delta p_n(x')}{dx'}$$

---

[7] In nondegenerate semiconductors, the diffusion coefficient D and the drift mobility μ for a given type of carrier are related by $D/\mu = kT/e$ so that we can easily calculate D from μ. The proof is not difficult; see, for example, S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd Edition (McGraw-Hill, 2006), Ch. 5.

[8] The intuitive derivation here for Eq. (3.5.10) was actually based on knowing the outcome from the continuity equation and using the fact that we have a very long *n*-side. In the latter case *only*, the oversimplified intuitive arguments above do indeed lead to Eq. (3.5.10); but not in a general treatment. One must also consider the change in the diffusion flux (the current). (See, for example, S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd Edition, McGraw-Hill, 2006), Ch. 6.

[9] The hole diffusion flux is $-D_h(dp/dx) = -D_h(d\Delta p/dx)$ and the diffusing charge is $+e$.

i.e.,

$$J_{D,\text{hole}} = \left(\frac{eD_h}{L_h}\right)\Delta p_n(0)\exp\left(-\frac{x'}{L_h}\right) \tag{3.5.11}$$

The above equation shows that hole diffusion current depends on location x' and decreases due to recombination. The total current at any location, however, is the sum of hole and electron contributions. The total current is independent of x' as indicated in Figure 3.17. The decrease in the minority carrier diffusion current with x' is made up by the increase in the current due to the drift of the majority carriers as schematically shown in Figure 3.17. The field in the neutral region is not totally zero but a small value, just sufficient to drift the huge number of majority carriers there and maintain a constant current.

We can now use the law of the junction to substitute for $\Delta p_n(0)$ in Eq. (3.5.11) in terms of the applied voltage V in Eq. (3.5.8). Further, we can eliminate p_{no} by $p_{no} = n_i^2/n_{no} = n_i^2/N_d$. Thus, at $x' = 0$, just outside the depletion region, from Eq. (3.5.11) the hole diffusion current density is

Hole diffusion current

$$J_{D,\text{hole}} = \left(\frac{eD_h n_i^2}{L_h N_d}\right)\left[\exp\left(\frac{eV}{k_B T}\right) - 1\right] \tag{3.5.12}$$

There is a similar expression for the electron diffusion current density $J_{D,\text{elec}}$ in the p-region. We will assume that the electron and hole currents do not change across the depletion region because, in general, the width of this region is quite narrow (and, for the time being, we neglect the recombination in the SCL). The electron current at $x = -W_p$ is the same as that at $x = W_n$. The total current density is then simply given by $J_{D,\text{hole}} + J_{D,\text{elec}}$, that is,

Shockley long diode equation

$$J = \left(\frac{eD_h}{L_h N_d} + \frac{eD_e}{L_e N_a}\right)n_i^2\left[\exp\left(\frac{eV}{k_B T}\right) - 1\right] \tag{3.5.13a}$$

or

Shockley long diode equation

$$J = J_{so}\left[\exp\left(\frac{eV}{k_B T}\right) - 1\right] \tag{3.5.13b}$$

This is the familiar diode equation with $J_{so} = \left[(eD_h/L_h N_d) + (eD_e/L_e N_a)\right]n_i^2$. It is frequently called the **Shockley equation**. It represents the *diffusion of minority carriers* in the neutral regions. The constant J_{so} depends not only on the doping, N_d, N_a, but also on the material

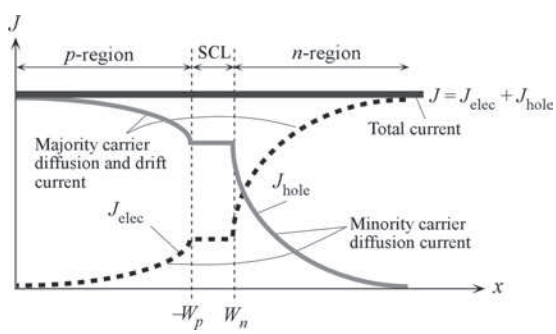


FIGURE 3.17 The total current anywhere in the device is constant. Just outside the depletion region it is primarily due to the diffusion of minority carriers and near the contacts it is primarily due to the drift of majority carriers.

via n_i, D_h, D_e, L_h, L_e. It is known as the **reverse saturation current density**, because if we apply a reverse bias $V = -V_r$ greater than the thermal voltage k_BT/e (25 mV), Eq. (3.5.13b) becomes $J = -J_{so}$.

We note that since the applied voltage drops across the depletion region, the built-in field is reduced by the applied field imposed by the bias, that is, the field at the junction becomes $E_o - E$, where E is the applied field at the junction. Because the field is reduced, we need less charges on either side of the junction. This can only be accommodated if the depletion regions shrink in width to W_p and W_n on the p- and n-sides, and the new total width $W = W_p + W_n$ is less than W_o. Equation (3.5.1) remains valid in the form $N_aW_p = N_dW_n$.

The *pn* junction in which the neutral regions l_n and l_p are shorter than the diffusion lengths L_h and L_n, respectively, is called a **short diode**. Its treatment is much easier than the long diode. There is essentially no time for the injected hole (minority carrier) in the n-side to recombine because it quickly diffuses and reaches the collection (negative) electrode, which is very close in distance. If we are not losing holes by recombination, then there is no reason for the hole diffusion current to decay along x' in this region as in the long diode where recombination is present. The hole diffusion current would remain constant if the hole concentration gradient is constant, that is, the excess hole concentration profile decreases linearly as shown in Figure 3.18. Thus, the hole diffusion current density $J_{D,\text{hole}}$ on the n-side is

$$J_{D,\text{hole}} = -eD_h\frac{d\Delta p_n(x')}{dx'} = eD_h\frac{p_n(0) - p_{no}}{l_n} = eD_h\frac{p_{no}e^{eV/k_BT} - p_{no}}{l_n}$$

so that

$$J_{D,\text{hole}} = \frac{eD_hn_i^2}{l_nN_d}\left[\exp\left(\frac{eV}{k_BT}\right) - 1\right]$$

A similar argument applies for $J_{\text{D,elec}}$ on the p-side so that the total current density is then

$$J = \left(\frac{eD_h}{l_nN_d} + \frac{eD_e}{l_pN_a}\right)n_i^2\left[\exp\left(\frac{eV}{k_BT}\right) - 1\right] \qquad \textbf{(3.5.14)}$$

Shockley short diode equation

Equation (3.5.14) is identical to Eq. (3.5.13a) with diffusion lengths L_h and L_e replaced by the widths of the neutral regions l_n and l_p, respectively. The short diode is a good example of a

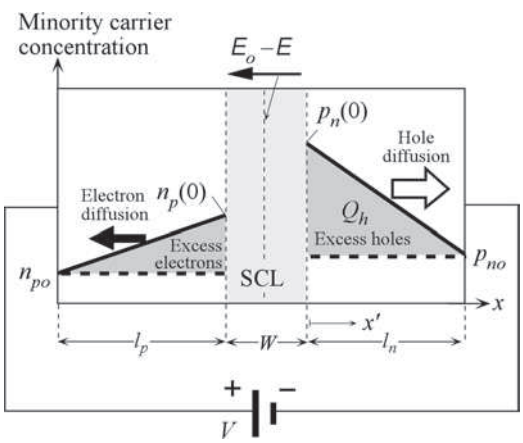


FIGURE 3.18 Minority carrier injection and diffusion in a short diode. Minority recombination in the neutral regions is negligible. $p_n(0)$ is the minority carrier concentration just outside the depletion region, and is controlled by the law of the junction, $p_n(0) = p_{no}\exp(eV/k_BT)$.

device in which the current is almost entirely due to the *diffusion of minority carriers*, and highlights the significant role played by diffusion.

C. Minority Carrier Charge Stored in Forward Bias

So far we have considered a *pn* junction with heavier doping on the *p*-side. The injected minority carriers, holes in this case, represent an *injected excess minority carrier charge*, Q_h, in the neutral region as shown in Figures 3.16 (a) and 3.18. There is, of course, also excess majority carrier charge so the region is neutral; but the excess majority carrier concentration (about the same as $p_n(0)$) is very small compared with the actual majority carrier concentration ($n_{no} = N_d$). For the long diode, the excess holes in the neutral *n*-side diffuse and recombine after τ_h seconds—the minority carrier lifetime. The charge Q_h should therefore disappear after a time τ_h. It must therefore be replenished at a rate Q_h/τ_h if Q_h is to be maintained at steady state DC operation. The external current supplies holes at exactly this rate to maintain the steady state DC conduction. Thus, $I = Q_h/\tau_h$.

In the case of a short diode, a hole that is injected into the *n*-side must diffuse across the width l_n of this neutral region to reach the negative electrode. The time required for holes to diffuse across the width l_n of the neutral *n*-region is called the **transit time** τ_t, or the **diffusion time**. From random walk examples in various materials textbooks,[10] τ_t is given by $\tau_t = l_n^2/2D_h$. To maintain DC operating conditions, the current must replenish holes exactly at a rate Q_h/τ_t to make up for those holes reaching the end of the *n*-side, that is, $I = Q_h/\tau_t$. We can now generalize the above discussion as follows:

DC currents

$$I = Q/\tau \text{ for a } \textbf{long diode} \quad \text{and} \quad I = Q/\tau_t \text{ for a } \textbf{short diode} \tag{3.5.15}$$

where τ is the **effective minority carrier lifetime**, $1/\tau = 1/\tau_h + 1/\tau_e$, and τ_t is the **effective transit time**, that is, $1/\tau_t = 1/\tau_{th} + 1/\tau_{te}$; subscripts refer to holes and electrons in the neutral *n*- and *p*-sides, respectively. Equation (3.5.15) applies under DC conditions only.

D. Recombination Current and the Total Current

So far we have assumed that, under a forward bias, the minority carriers diffusing and recombining in the neutral regions are supplied by the external current. However, some of the minority carriers will recombine in the depletion region. The external current must therefore also supply the carriers lost in the recombination process in the SCL, as illustrated in Figure 3.19 (a). Consider for simplicity a symmetrical *pn* junction as in Figure 3.19 (b) under forward bias. At the junction center *C*, the hole and electron concentrations are p_M and n_M and are equal. We can find the SCL recombination current by considering electrons recombining in the *p*-side in W_p and holes recombining in the *n*-side in W_n as shown by the shaded areas *ABC* and *BCD*, respectively, in Figure 3.19 (b). Suppose that the **mean hole recombination time** in W_n is τ_h and **mean electron recombination time** in W_p is τ_e. The rate at which the electrons in *ABC* are recombining is the area *ABC* (nearly all injected electrons) divided by τ_e. The electrons are replenished by the diode current. Similarly, the rate at which holes in *BCD* are recombining is the area *BCD* divided by τ_h. Thus, the recombination current density is

$$J_{\text{recom}} = \frac{eABC}{\tau_e} + \frac{eBCD}{\tau_h}$$

[10]The expression for the mean square distance (l^2) diffused in a time interval τ, $l^2 = 2D\tau$, is derived in S. O. Kasap, *Principles of Electronic Materials and Devices*, 3rd Edition (McGraw-Hill, 2006), Ch. 1.

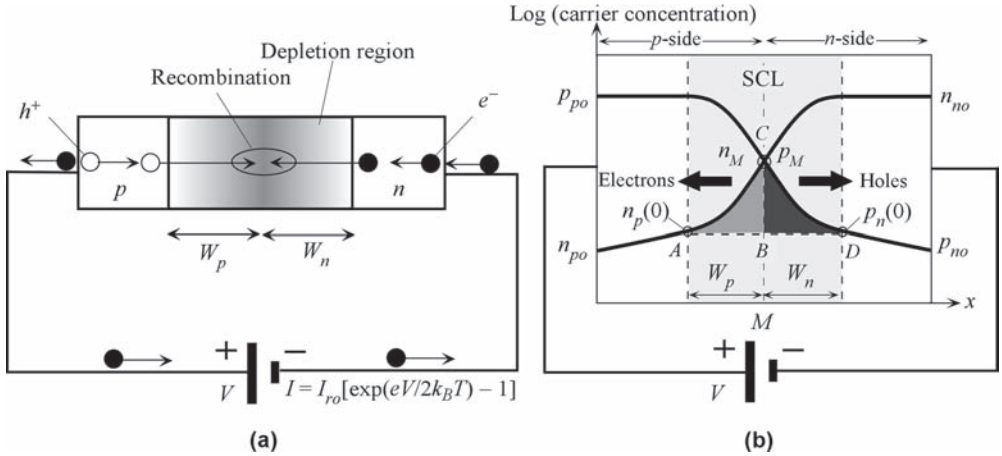


FIGURE 3.19 Forward-biased *pn* junction, the injection of carriers, and their recombination in the SCL. (a) Recombination of electrons and holes in the depletion region involves the current supplying the carriers. (b) A symmetrical *pn* junction for calculating the recombination current.

We can evaluate the areas *ABC* and *BCD* by taking them as triangles, $ABC \approx (1/2)W_p n_M$, *etc.*, so that

$$J_{\text{recom}} = \frac{e\frac{1}{2}W_p n_M}{\tau_e} + \frac{e\frac{1}{2}W_n p_M}{\tau_h} \tag{3.5.16}$$

Under steady state and equilibrium conditions, assuming a nondegenerate semiconductor, we can use Boltzmann statistics in Eq. (3.5.4) to relate these concentrations to the potential energy. At *A*, the hole potential energy is zero ($V = 0$) and at *M* it is $e(V_o - V)/2$ so that

$$\frac{p_M}{p_{po}} = \exp\left[-\frac{e(V_o - V)}{2k_BT}\right]$$

Since V_o depends on dopant concentrations and n_i as in Eq. (3.5.6) and further $p_{po} = N_a$, we can simplify the above to

$$p_M = n_i \exp\left(\frac{eV}{2k_BT}\right)$$

and there is a similar expression for n_M. We can substitute for p_M and n_M in Eq. (3.5.16) to find the recombination current for $V > k_BT/e$, that is,

$$J_{\text{recom}} = \frac{en_i}{2}\left(\frac{W_p}{\tau_e} + \frac{W_n}{\tau_h}\right)\exp\left(\frac{eV}{2k_BT}\right) \tag{3.5.17a}$$

Recombination current

From a better mathematical treatment, the expression for the recombination current can be shown to be[11]

$$J_{\text{recom}} = J_{ro}\left[\exp(eV/2k_BT) - 1\right] \tag{3.5.17b}$$

SCL recombination current

where J_{ro} is the pre-exponential constant in Eq. (3.5.17a).

[11] This is generally proved in advanced texts.

Equation (3.5.17) is the current that supplies the carriers that recombine in the depletion region. It does not stipulate what the recombination process is, but only the carriers recombine in this region. The recombination could be direct (as in high quality GaAs *pn* junctions) or indirect through defects or impurities in the SCL.

In many III–V *compound* semiconductors, the carriers also recombine at crystal surfaces and interfaces where, usually, there are defects that encourage recombination; or act as recombination centers. The external current also has to supply carriers that are lost through surface recombination. The exact *I–V* dependence for surface recombination is difficult to derive but has a voltage dependence quite similar to recombination in the SCL in Eq. (3.5.17b). Thus, surface recombination results in a current that can be described by

Surface recombination current

$$I_{\text{surf}} = I_{sro}\left[\exp\left(eV/2k_BT\right) - 1\right] \tag{3.5.18}$$

where I_{sro} is the pre-exponential constant that depends on the surface recombination process, and the surface area involved in the recombination.

The total current into the diode will supply carriers for minority carrier diffusion in the neutral regions and recombination in the space charge layer and at the surfaces and interfaces so that it will *be the sum* of Eqs. (3.5.13b), (3.5.17b), and (3.5.18). Generally the diode current is written as

General diode equation

$$I = I_o\left[\exp\left(\frac{eV}{\eta k_BT}\right) - 1\right] \tag{3.5.19}$$

where I_o is a constant and η, called the **diode ideality factor**, is 1 for diffusion-controlled and 2 for recombination-controlled forward current characteristics.

Figure 3.20 shows typical semilogarithmic plots of the forward *I–V* characteristics of three types of *pn* junctions based on Ge, Si, and GaAs. The Ge *pn* junction follows the Shockley model with $\eta = 1$ and the current is due to diffusion in the neutral regions. Both Si and GaAs diodes initially follow the recombination model and then the Shockley model. Many Si and GaAs diodes have an η that is between 1 and 2 for the *I–V* ranges of interest.

Notice that, at a given current, say 1 mA, the voltage across a Ge diode is about 0.2 V, it is 0.7 V across the Si diode, and almost 1.1 V across the GaAs diode. The increase in the turn-on

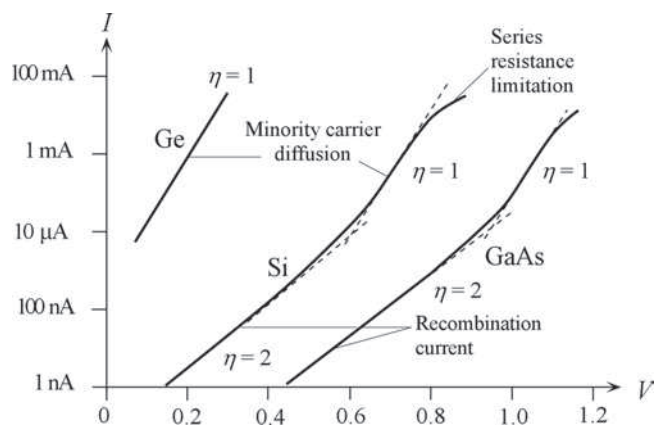


FIGURE 3.20 Schematic representation of the forward *I–V* characteristics of three typical *pn* junctions based on Ge, Si, and GaAs.

voltage that results in a significant current from 0.2 V to 1.1 V follows the trend for the increase in the bandgap E_g from Ge (0.66 eV) to Si (1.12 eV) to GaAs (1.42 eV). The pre-exponential factor I_o in the diode current equation contains n_i, which depends exponentially on the bandgap E_g. As E_g increases, n_i falls sharply by Eq. (3.2.6), and results in very small I_o values. Thus, a larger applied voltage is needed to bring I_o up to any significant current.

It is clear that, under a forward bias, the *p*-side acts to inject holes into the *n*-side, and the *n*-side acts to inject electrons into the *p*-side. The latter concept will become useful in understanding the operation of heterostructures discussed below.

3.6 *pn* JUNCTION REVERSE CURRENT

When a *pn* junction is reverse biased, the reverse current is typically very small. The reverse bias across a *pn* junction is illustrated in Figure 3.21 (a). The applied voltage drops mainly across the resistive depletion region, which becomes wider. The applied field is in the same direction as the built-in field and adds to it, so that the field at the junction becomes $E_o + E$, and large, where E is the field imposed by the applied forward bias. There is therefore a larger electric field inside the depletion region due to the reverse bias. To accommodate this larger field, the widths of the depletion region on the *p*- and the *n*-side widen to expose more ionized dopants as indicated in Figure 3.21 (a). The movement of electrons in the *n*-region toward the positive battery terminal cannot be sustained because there is no electron supply to this *n*-side. The *p*-side cannot supply electrons to the *n*-side because it has almost none. However, there is a small reverse current due to two causes.

The applied voltage increases the built-in potential barrier, as illustrated in Figure 3.21 (b). The electric field in the SCL is larger than the built-in internal field E_o. The small number of holes on the *n*-side near the depletion region become extracted and swept by the field across the SCL over to the *p*-side. This small current can be maintained by the diffusion of holes from the *n*-side bulk to the SCL boundary. Assume that the reverse bias $V_r > 25\text{ mV} = k_BT/e$. The hole concentration $p_n(0)$ just outside the SCL is nearly zero by the law of the junction, Eq. (3.5.8), whereas the hole concentration in the bulk (or near the negative terminal) is the equilibrium

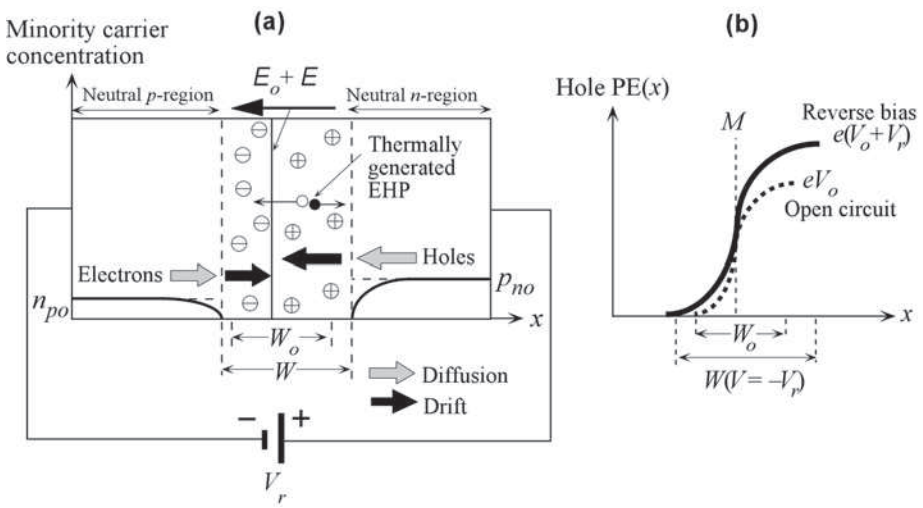


FIGURE 3.21 Reverse-biased *pn* junction. (a) Minority carrier profiles and the origin of the reverse current. (b) Hole PE across the junction under reverse bias.

concentration p_{no}, which is small. There is therefore a small concentration gradient and hence a small hole diffusion current toward the SCL as shown in Figure 3.21 (a). Similarly, there is a small electron diffusion current from the bulk p-side to the SCL. Within the SCL, these carriers are drifted by the field. This minority carrier diffusion is accounted in the Shockley model with $V = -V_r$. The reverse current is given by Eq. (3.5.13) with a negative voltage which leads to a diode current density of $-J_{so}$ called the **reverse saturation current density**. The value of J_{so} depends only on the material via n_i, μ_h, μ_e, the dopant concentrations, *etc.*, but not on the voltage ($V_r > k_BT/e$). Furthermore, as J_{so} depends on n_i^2, it is strongly temperature dependent. In some books, it is stated that the thermal generation of minority carriers in the neutral region within a diffusion length to the SCL, the diffusion of these carriers to the SCL, and their subsequent drift through the SCL is the cause of the reverse current. This description, in essence, is identical to the Shockley model, and the description above.

The thermal generation of electron–hole pairs in the SCL, as shown in Figure 3.21 (a), can also contribute to the observed reverse current since the internal field in this layer will separate the electron and hole, and drift them toward the neutral regions. This drift will result in an external current in addition to the reverse current due to the diffusion of minority carriers. The theoretical evaluation of SCL generation current involves an in-depth knowledge of the charge carrier generation processes via recombination centers, which is discussed in advanced texts. Suppose that τ_g is the **mean time to generate an electron–hole pair** by virtue of thermal vibrations of the lattice; τ_g is also called the **mean thermal generation time**. Given τ_g, the rate of thermal generation per unit volume must be n_i/τ_g because it takes on average τ_g seconds to create n_i number of EHPs per unit volume. Furthermore, since WA, where A is the cross-sectional area, is the volume of the depletion region, the rate of EHP, or charge carrier, generation is $(AWn_i)/\tau_g$. Both holes and electrons drift in the depletion region and both contribute to the current. The observed current density must be $e(Wn_i)/\tau_g$. Therefore, the reverse current density component due to thermal generation of electron–hole pairs within the SCL should be given by

EHP thermal generation in SCL

$$J_{\text{gen}} = \frac{eWn_i}{\tau_g} \tag{3.6.1}$$

The total reverse current density J_{rev} is the sum of the diffusion and generation components, that is,

Total reverse current

$$J_{\text{rev}} = \left(\frac{eD_h}{L_hN_d} + \frac{eD_e}{L_eN_a}\right)n_i^2 + \frac{eWn_i}{\tau_g} \tag{3.6.2}$$

which is shown schematically in Figure 3.22 (a). The thermal generation component J_{gen} in Eq. (3.6.1) increases with the reverse bias V_r because the SCL width W widens with V_r.

The terms in the reverse current in Eq. (3.6.2) are predominantly controlled by n_i^2 and n_i. Their relative importance depends not only on the semiconductor properties but also on the temperature since $n_i \propto \exp(-E_g/2k_BT)$. Figure 3.22 (b) shows the reverse current I_{rev} in the dark in a Ge pn junction plotted as $\ln(I_{\text{rev}})$ vs. $1/T$ to highlight the two different processes in Eq. (3.6.2). The measurements in Figure 3.22 (b) show that above 238 K, I_{rev} is controlled by n_i^2 because the slope of $\ln(I_{\text{rev}})$ vs. $1/T$ yields an E_g of approximately 0.63 eV, close to the expected E_g of about 0.66 eV in Ge. Below 238 K, I_{rev} is controlled by n_i because the slope of $\ln(I_{\text{rev}})$ vs. $1/T$ is

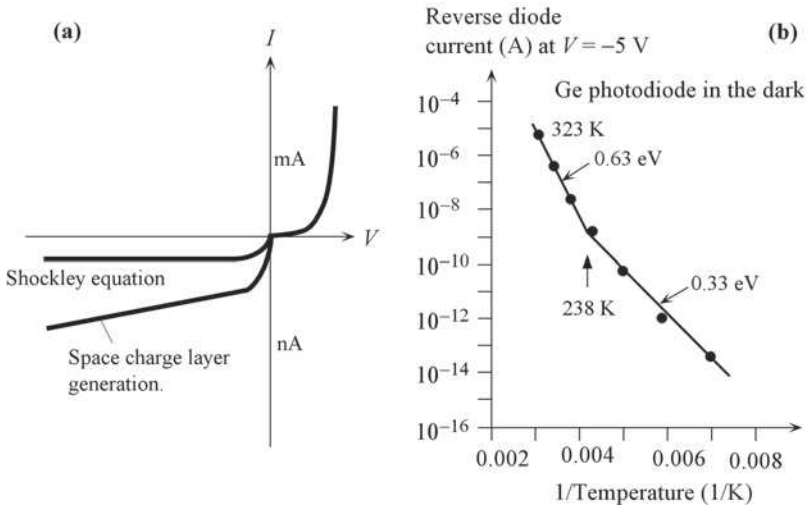


FIGURE 3.22 (a) Schematic illustration of the reverse current of a *pn* junction and the two contributing factors: diffusion and thermal generation. (The *I*-axis changes unit for forward and reverse currents.) (b) Reverse diode current in a Ge *pn* junction as a function of temperature in a $\ln(I_{rev})$ vs. $1/T$ plot. Above 238 K, I_{rev} is controlled by n_i^2, and below 238 K it is controlled by n_i. The vertical axis is a logarithmic scale with actual current values. (*Source:* Data plotted from D. Scansen and S. O. Kasap, *Cnd. J. Physics., 70,* 1070, 1992.)

equivalent to an $E_g/2$ of approximately 0.33 eV. In this range, the reverse current is due to EHP generation in the SCL via defects and impurities (recombination centers).

3.7 *pn* JUNCTION DYNAMIC RESISTANCE AND CAPACITANCES

A. Depletion Layer Capacitance

It is apparent that the depletion region or space charge layer of a *pn* junction has positive charges in W_n and negative charges in W_p, which have been separated from each other, similar to a parallel plate capacitor as indicated in Figures 3.23 (a) and 3.15 (d). We also know that an applied voltage to the *pn* junction modifies the width W of the SCL; it increases with the reverse bias. If A is the cross-sectional area, the stored charge in the depletion region is $+Q = eN_dW_nA$ on the n-side and $-Q = -eN_aW_pA$ on the p-side. Unlike in the case of a parallel plate capacitor, Q does not depend linearly on the voltage V across the device.

It is useful to define an **incremental capacitance** which relates the incremental change in the charge stored in the depletion region to an incremental voltage change across the *pn* junction. When the voltage V across a *pn* junction changes by dV to $V + dV$, then W also changes and, as a result, the amount of charge in the depletion region becomes $Q + dQ$. The **depletion layer capacitance** C_{dep} is defined by[12]

$$C_{dep} = \left|\frac{dQ}{dV}\right| \qquad \textbf{(3.7.1)}$$

Depletion layer capacitance

[12]This is actually an *incremental* capacitance; this is implied by the definition in Eq. (3.7.1).

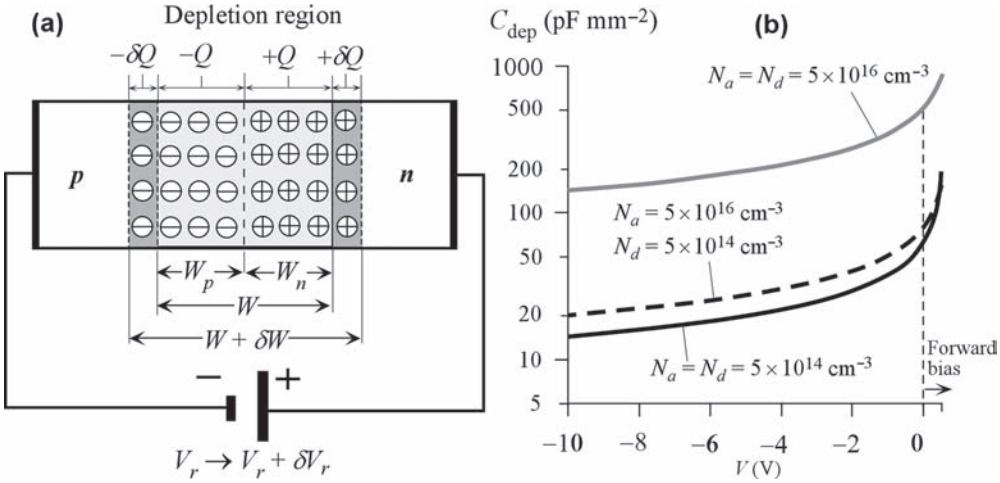


FIGURE 3.23 (a) Depletion region has negative ($-Q$) charges in W_p and positive ($+Q$) charges in W_n, which are separated as in a capacitor. Under a reverse bias V_r, the charge on the n-side is $+Q$. When the reverse bias is increased by δV_r, the charge Q increases by δQ. (b) C_{dep} vs. voltage across an abrupt pn junction for three different sets of dopant concentrations. (Note that the vertical scale is logarithmic.)

If the applied voltage is V then the voltage across the depletion layer W is $V_o - V$ and the depletion region width is

SCL width and voltage

$$W = \left[\frac{2\varepsilon(N_a + N_d)(V_o - V)}{eN_aN_d}\right]^{1/2} \tag{3.7.2}$$

The amount of charge (on any one side of the depletion layer) is $|Q| = eN_dW_nA = eN_aW_pA$ and $W = W_n + W_p$. We can therefore substitute for W in Eq. (3.7.2) in terms of Q and then differentiate it to obtain dQ/dV. The final result for the depletion capacitance is

Depletion layer capacitance

$$C_{dep} = \frac{\varepsilon A}{W} = \frac{A}{(V_o - V)^{1/2}}\left[\frac{e\varepsilon(N_aN_d)}{2(N_a + N_d)}\right]^{1/2} \tag{3.7.3}$$

We should note that C_{dep} is given by the same expression as that for the parallel plate capacitor, $\varepsilon A/W$, but with W being voltage dependent by virtue of Eq. (3.7.2). Putting a reverse bias $V = -V_r$ in Eq. (3.7.3) shows that C_{dep} decreases with increasing V_r in a $C_{dep} \propto 1/V_r^{1/2}$ fashion. The reason is that the reverse bias widens W and hence increases the average separation of $+Q$ and $-Q$. Typically C_{dep} under reverse bias is in the 10–10^3 pF mm^{-2} depending on N_a and N_d as shown in Figure 3.23 (b).

If $N_a \gg N_d$ as in the case of a p^+n junction, then

Depletion layer capacitance for p^+n

$$C_{dep} = A\left[\frac{e\varepsilon N_d}{2(V_o - V)}\right]^{1/2} \tag{3.7.4}$$

which does not depend on the p-side doping. C_{dep} is present in both forward- and reverse-biased junction.

B. Dynamic Resistance and Diffusion Capacitance for Small Signals

Consider a long *pn* junction diode in which the *p*-side is heavily doped. Under forward bias, minority carriers are injected and diffuse through the neutral regions as shown in Figure 3.16 (a). The excess hole concentration profile $\Delta p_n(x)$ in the *n*-side due the applied forward bias is shown in Figure 3.24 (a). The injected charge Q is the area under the excess hole concentration profile $\Delta p_n(x)$. It is called the stored minority carrier charge, which is due to diffusing holes in the *n*-side. Recall that the hole concentration profile is a steady state profile. If the hole lifetime is τ_h, the recombination rate of the holes in the *n*-side is determined by Q/τ_h. The current must replenish the holes lost exactly at this rate to maintain a steady state profile. Thus, $I = Q/\tau_h$.

Suppose we increase the diode voltage from V to $V + \delta V$. (This can also be achieved by increasing the drive current from I to $I + \delta I$, which is normally the case in practice.) More holes will be injected into the *n*-side as a result of the law of the junction as shown in Figure 3.24 (a). A change in the diode voltage V to $V + \delta V$, as shown in Figure 3.24 (b), results in change in the forward current. For small changes, we can relate the two because the diode current $I = I_o \exp(eV/k_BT)$, which can be differentiated with respect to V. We define the **dynamic resistance**, also known as the **incremental resistance**, of the diode as

$$r_d = \frac{dV}{dI} = \frac{V_{\text{th}}}{I} \qquad \textbf{(3.7.5)}$$

Dynamic resistance

where $V_{\text{th}} = k_BT/e$ is the thermal voltage (26 mV); r_d can be quite small, 2.5 Ω at $I = 10$ mA. The dynamic resistance r_d does not represent a true resistance for heat dissipation, but rather a relationship between current and voltage increments imposed by the *pn*-junction behavior. It represents the *diode action* as embedded in $I = I_o \exp(eV/k_BT)$.

When the diode voltage is increased from V to $V + \delta V$, the excess hole concentration just inside the *n*-side will increase from $\Delta p_n(0)$ to $\Delta p_n'(0)$. Additional charge δQ is injected into the *n*-side as shown in Figure 3.24 (a). The increase in the minority carrier charge Q due to the increase

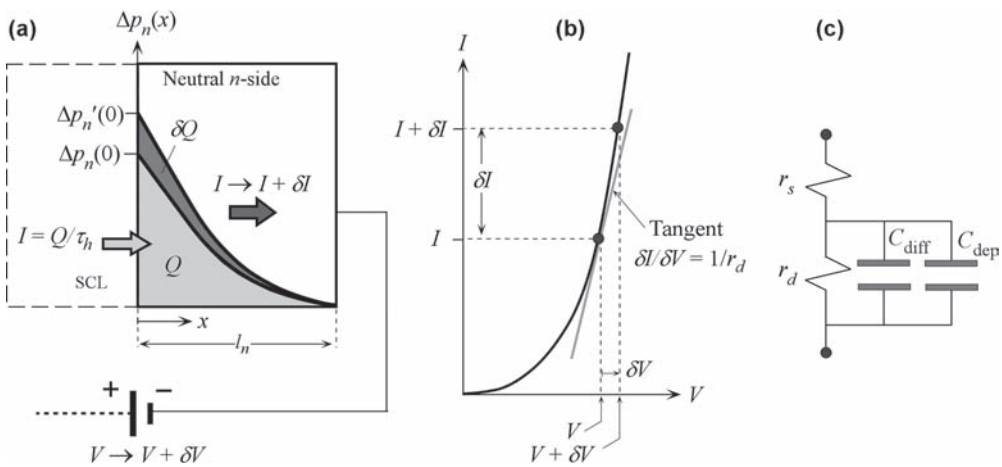


FIGURE 3.24 (a) The forward voltage across a *pn* junction increases by δV, which leads to further minority carrier injection and a larger forward current, which increases by δI. Additional minority carrier charge δQ is injected into the *n*-side. The increase δQ in charge stored in the *n*-side with δV appears as if there is a capacitance across the diode. (b) The increase δV results in an increase δI in the diode current. The dynamic or incremental resistance $r_d = \delta V/\delta I$. (c) A simplified equivalent circuit for a forward-biased *pn* junction for small signals.

in the voltage across the *pn* junction represents a *capacitive behavior* in the sense it takes time to store the additional charge δQ. The additional stored charge δQ has to be removed when $V + \delta V$ returns back to the original value V. The excess charge δQ disappears by recombination, which takes time; again capacitance-like behavior. The **diffusion** or **storage capacitance** is defined as $C_{\text{diff}} = dQ/dV$. We can easily find C_{diff}. Given $I = Q/\tau_h$ (τ_h being the minority carrier lifetime) under forward bias as in Eq. (3.5.15), we can differentiate Q with respect to V to find

Diffusion capacitance

$$C_{\text{diff}} = \frac{\tau_h I}{V_{\text{th}}} = \frac{\tau_h}{r_d} \tag{3.7.6}$$

It should be mentioned that although under forward bias, C_{dep} can be significant (a few hundred pF mm^{-2}) the diffusion or storage capacitance, C_{diff}, due to the injection minority carriers that are diffusing in the neutral regions far exceeds C_{dep}. In fact, C_{diff} can be several nF mm^{-2}, and a limiting factor in the speed of *pn* junction devices operating in the forward bias regime. We can easily develop a small signal equivalent circuit for a forward-biased junction carrying a current I. There will be a dynamic resistance r_d with two capacitances C_{diff} and C_{dep} in parallel. We should also include a small resistance r_s in series as in Figure 3.24 (c) to represent the resistance of the neutral bulk semiconductor regions. The current has to flow through r_s, which gives rise to a voltage drop through the bulk semiconductor so that not all of applied V drops across the depletion region. Instead of V_o being reduced by V, it is reduced by $V - Ir_s$.

It is important to note that Eq. (3.7.6) is grossly simplified. A proper analysis should impose an AC signal of the form $\delta V = V_m \cos(\omega t)$ on top of the DC operating conditions, where ω is the angular frequency of the modulating signal, and derive the signal current δI, and hence the admittance.[13] The final result makes r_d as above in Eq. (3.7.5), but C_{diff} is different by a factor 2, that is,

Diffusion capacitance, long diode (AC)

$$C_{\text{diff}} = \frac{\tau_h I}{2V_{\text{th}}} = \frac{\tau_h}{2r_d} \tag{3.7.7}$$

The lifetime τ_h is the minority carrier lifetime in the lesser doped region (holes in the *n*-side). Further Eq. (3.7.7) is valid when $\omega < 1/\tau_h$. At high frequencies ($\omega > 1/\tau_h$), both r_d and C_{diff} become frequency dependent. For a short diode, r_d remains the same. The diffusion capacitance, however, becomes

Diffusion capacitance, short diode (AC)

$$C_{\text{diff}} = \frac{\tau_t I}{V_{\text{th}}} = \frac{\tau_t}{r_d} \tag{3.7.8}$$

where τ_t is the minority carrier transit (diffusion) time. If we have a short diode as in Figure 3.18, and we decrease the voltage, the excess holes will have to diffuse out. The time it takes for a hole to diffuse across the neutral region width l_n is τ_t, and is given by $\tau_t = l_n^2/2D_h$. For a short diode, $C_{\text{diff}} = \tau_t/r_d$.

3.8 RECOMBINATION LIFETIME

A. Direct Recombination

Consider the recombination of an electron and hole in a direct bandgap semiconductor, for example, doped GaAs. Recombination involves a direct meeting of an electron and a hole. Suppose that excess electrons and holes have been injected, as would be in a *pn*-junction under

[13] J. Gower, *Optical Communications Systems*, 2nd Edition (Prentice Hall, Pearson Education, 1993), Ch. 15.

forward bias, and that Δn_p is the excess electron concentration and Δp_p is the excess hole concentration in the *neutral p*-side of a GaAs *pn* junction. Injected electron and hole concentrations would be the same to maintain charge neutrality, that is, $\Delta n_p = \Delta p_p$. Thus, at any instant

$$n_p = n_{po} + \Delta n_p = \text{instantaneous } \textit{minority} \text{ carrier concentration}$$

and

$$p_p = p_{po} + \Delta n_p = \text{instantaneous } \textit{majority} \text{ carrier concentration}$$

The instantaneous recombination rate will be proportional to both the electron and hole concentrations at that instant, that is, $n_p p_p$. Suppose that the thermal generation rate of EHPs is G_{thermal}. The net rate of *change* of Δn_p is[14]

$$d\Delta n_p/dt = -Bn_p p_p + G_{\text{thermal}} \tag{3.8.1}$$

where B is called the **direct recombination capture coefficient**. In equilibrium $d\Delta n_p/dt = 0$ so that setting Eq. (3.8.1) to zero and using $n_p = n_{po}$ and $p_p = p_{po}$, where the subscript o refers to thermal equilibrium concentrations, we find $G_{\text{thermal}} = Bn_{po}p_{po}$. Thus, the rate of change in Δn_p is

Excess carrier rate of change due to recombination

$$d\Delta n_p/dt = -B(n_p p_p - n_{po}p_{po}) \tag{3.8.2}$$

In many instances the rate of change $d\Delta n_p/dt$ is proportional to Δn_p. It is therefore useful to define an **excess minority carrier recombination time** (**lifetime**) τ_e by

Recombination time definition

$$d\Delta n_p/dt = -\Delta n_p/\tau_e \tag{3.8.3}$$

Consider two practical cases where injected excess minority carrier concentration Δn_p is either much less than the actual equilibrium majority carrier concentration p_{po}, or greater. The two cases correspond to weak and strong injection based on Δn_p compared with p_{po}.

In **weak injection**, $\Delta n_p \ll p_{po}$. Then $n_p \approx \Delta n_p$ and $p_p \approx p_{po} + \Delta p_p \approx p_{po} = N_a$, that is, the acceptor concentration. Therefore with these approximations in Eq. (3.8.2) we obtain

$$d\Delta n_p/dt = -BN_a\Delta n_p \tag{3.8.4}$$

Thus, comparing with Eq. (3.8.3), we find the lifetime to be

Weak injection lifetime

$$\tau_e = 1/BN_a \tag{3.8.5}$$

and is constant under weak injection conditions as here.

In **strong injection**, $\Delta n_p \gg p_{po}$. It is easy to show that with this condition, Eq. (3.8.2), becomes

$$d\Delta n_p/dt = -B\Delta p_p\Delta n_p = B(\Delta n_p)^2 \tag{3.8.6}$$

which gives

Strong injection lifetime

$$\tau_e = 1/B\Delta n_p = 1/Bn_p \tag{3.8.7}$$

[14] Normally partial derivatives $\partial/\partial t$ would be used since the excess concentration can also depend on x. We ignore this dependence for now to derive the expressions we need for the recombination lifetime.

so that under high-level injection conditions, the lifetime τ_e is inversely proportional to the injected carrier concentration. When a light-emitting diode (LED) is modulated under high injection levels, for example, the lifetime of the minority carriers is therefore not constant, which in turn leads to the distortion of the modulated light output.

B. Indirect Recombination

Consider the recombination of minority carriers such as electrons in an extrinsic indirect bandgap semiconductor such as Si or Ge that have been doped *p*-type. In an indirect bandgap semiconductor, the recombination mechanism involves a recombination center at an energy E_r in Figure 3.14 (c). Such centers are usually crystal defects or impurities in the crystal. The center acts as a third body in the electron and hole recombination process to satisfy the requirements of conservation of momentum. We can view the recombination process as follows. Recombination occurs when an electron is captured by the recombination center at the energy level E_r. As soon as the electron is captured, it will recombine with a hole because holes are abundant in a *p*-type semiconductor. In other words, since there are many majority carriers, the limitation on the rate of recombination is the actual capture of the minority carrier by the center. Suppose that τ_e is the electron recombination time or lifetime, an average time it takes for an electron to recombine. Then, $1/\tau_e$ is the rate of capture per electron. Since the injected electrons will have to be captured by the centers, $1/\tau_e$ is proportional to the capture (recombination) cross-section S_r of the center, the concentration of centers N_r, and also the mean speed of the electron v_{th} (approximately its thermal velocity) all of which increase the rate of capture. Thus,

Indirect recombination via centers

$$\tau_e = \frac{1}{S_r N_r v_{\text{th}}} \tag{3.8.8}$$

where Eq. (3.8.8) is valid under small injection conditions, that is, $n_p < p_{po}$. There is a more general treatment of indirect recombination called the Shockley-Read-Hall model of indirect recombination and generation, which is treated in more advanced semiconductor physics textbooks; that theory eventually arrives at Eq. (3.8.8) for low-level injection conditions.

EXAMPLE 3.8.1 A direct bandgap *pn* junction

Consider a symmetrical GaAs *pn* junction which has the following properties. N_a (*p*-side doping) = N_d (*n*-side doping) = $10^{17}\,\text{cm}^{-3}$ (or $10^{23}\,\text{m}^{-3}$), direct recombination coefficient[15] $B \approx 2 \times 10^{-16}\,\text{m}^3\,\text{s}^{-1}$, cross-sectional area $A = 1\,\text{mm}^2$. Suppose that the forward voltage V across the diode is 0.80 V. What is the diode current due to minority carrier diffusion at 27°C (300 K) assuming direct recombination? If the mean minority carrier lifetime in the depletion region were to be the same as this lifetime, what would be the recombination component of the diode current? What are the two contributions at $V = 1.05$ V? What is your conclusion?

Note that at 300 K, GaAs has an intrinsic concentration (n_i) of $2.1 \times 10^6\,\text{cm}^{-3}$ and a relative permittivity (ε_r) of 13.0. The hole drift mobility (μ_h) in the *n*-side is $250\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ and electron drift mobility (μ_e) in the *p*-side is 5000 $\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ (these are at the doping levels given).

[15] The direct recombination coefficient B at room temperature is typically of the order $10^{10}\,\text{cm}^3\,\text{s}^{-1}$ for many III–V direct bandgap semiconductors; see, for example, E. F. Schubert, *Light-Emitting Diodes*, 2nd Edition (Cambridge University Press, 2006), Table 3.1.

Solution

Assuming weak injection, we can calculate the recombination times τ_e and τ_h for electrons and holes recombining in the neutral p- and n-regions, respectively. Using S.I. units throughout, taking $k_BT/e = 0.02585\,\text{V}$, and since this is a symmetric device,

$$\tau_h = \tau_e \approx \frac{1}{BN_a} = \frac{1}{(2.0 \times 10^{-16}\,\text{m}^3\,\text{s}^{-1})(1 \times 10^{23}\,\text{m}^{-3})} = 5.00 \times 10^{-8}\,\text{s} \quad \text{or} \quad 50.0\,\text{ns}$$

To find the Shockley current we need the diffusion coefficients and lengths. The Einstein relation[7] gives the diffusion coefficients as

$$D_h = \mu_h k_B T/e = (0.02585)(250 \times 10^{-4}) = 6.46 \times 10^{-4}\,\text{m}^2\,\text{s}^{-1}$$
$$D_e = \mu_e k_B T/e = (0.02585)(5000 \times 10^{-4}) = 1.29 \times 10^{-2}\,\text{m}^2\,\text{s}^{-1}$$

where k_BT/e was taken as 0.02585 V. The diffusion lengths are easily calculated as

$$L_h = (D_h\tau_h)^{1/2} = \left[(6.46 \times 10^{-4}\,\text{m}^2\,\text{s}^{-1})(50.0 \times 10^{-9}\,\text{s})\right]^{1/2} = 5.69 \times 10^{-6}\,\text{m}$$
$$L_e = (D_e\tau_e)^{1/2} = \left[(1.29 \times 10^{-2}\,\text{m}^2\,\text{s}^{-1})(50.0 \times 10^{-9}\,\text{s})\right]^{1/2} = 2.54 \times 10^{-5}\,\text{m}$$

Notice that the electrons diffuse much further in the p-side. The reverse saturation current due to diffusion in the neutral regions is

$$I_{so} = A\left(\frac{D_h}{L_hN_d} + \frac{D_e}{L_eN_a}\right)en_i^2$$
$$= (10^{-6})\left[\frac{6.46 \times 10^{-4}}{(5.69 \times 10^{-6})(10^{23})} + \frac{1.29 \times 10^{-2}}{(2.54 \times 10^{-5})(10^{23})}\right](1.6 \times 10^{-19})(2.1 \times 10^{12})^2$$
$$\approx 4.4 \times 10^{-21}\,\text{A}$$

Thus, the forward diffusion current at $V = 0.80$ V is

$$I_{\text{diff}} = I_{so}\exp\left(\frac{eV}{k_BT}\right)$$
$$= (4.4 \times 10^{-21}\,\text{A})\exp\left(\frac{0.80\,\text{V}}{0.02585\,\text{V}}\right) = 1.2 \times 10^{-7}\,\text{A} \quad \text{or} \quad 0.12\,\mu\text{A}$$

Recombination component of the current is quite difficult to calculate because we need to know the mean electron and hole recombination times in the SCL. Suppose that, as a first order, we assume that these recombination times are as above.

The built-in voltage V_o is

$$V_o = \frac{k_BT}{e}\ln\left(\frac{N_aN_d}{n_i^2}\right) = (0.02585)\ln\left[\frac{10^{23}10^{23}}{(2.1 \times 10^{12})^2}\right] = 1.27\,\text{V}$$

Depletion layer width W is

$$W = \left[\frac{2\varepsilon(N_a + N_d)(V_o - V)}{eN_aN_d}\right]^{1/2}$$
$$= \left[\frac{2(13)(8.85 \times 10^{-12}\,\text{F}\,\text{m}^{-1})(10^{23} + 10^{23}\,\text{m}^{-3})(1.27 - 0.80\,\text{V})}{(1.6 \times 10^{-19}\,\text{C})(10^{23}\,\text{m}^{-3})(10^{23}\,\text{m}^{-3})}\right]^{1/2}$$
$$= 1.16 \times 10^{-7}\,\text{m} \quad \text{or} \quad 0.116\,\mu\text{m}$$

As this is a symmetric diode, $W_p = W_n = W/2$. The pre-exponential I_{ro} is

$$I_{ro} = \frac{Aen_i}{2}\left(\frac{W_p}{\tau_e} + \frac{W_n}{\tau_h}\right) = \frac{Aen_i}{2}\left(\frac{W}{\tau_e}\right)$$
$$= \frac{(10^{-6})(1.6 \times 10^{-19})(2.1 \times 10^{12})}{2}\left(\frac{1.16 \times 10^{-7}}{5.00 \times 10^{-8}}\right) \approx 3.9 \times 10^{-13}\,\text{A}$$

so that at $V = 0.80$ V,

$$I_{\text{recom}} \approx I_{ro}\exp\left(\frac{eV}{2k_BT}\right)$$
$$\approx (3.9 \times 10^{-13}\,\text{A})\exp\left[\frac{0.8\ \text{V}}{2(0.02585\ \text{V})}\right] = 2.0 \times 10^{-6}\,\text{A} \quad \text{or} \quad 2.0\,\mu\text{A}$$

The recombination current is more than an order of magnitude greater than the diffusion current. If we repeat the calculation for a voltage of 1.05 V across the device, then we would find $I_{\text{diff}} = 1.9$ mA and $I_{\text{recom}} = 0.18$ mA, where I_{diff} dominates the current. Thus, as the voltage increases across a GaAs *pn* junction, the ideality factor η is initially 2 but then becomes 1 as shown in Figure 3.20.

The EHP recombination that occurs in the SCL and the neutral regions in this GaAs *pn* junction case would result in photon emission, with a photon energy that is approximately E_g. This direct recombination of injected minority carriers and the resulting emission of photons represent the principle of operation of the light-emitting diode (LED).

3.9 *pn* JUNCTION BAND DIAGRAM

A. Open Circuit

Consider a *p*-type and an *n*-type semiconductor from the same material (*e.g.*, Si) that are isolated from each other as in Figure 3.25 (a). The vacuum level, where the electron is free and has zero potential energy, is also shown and is common to both. The bandgap E_g and the electron affinity χ are

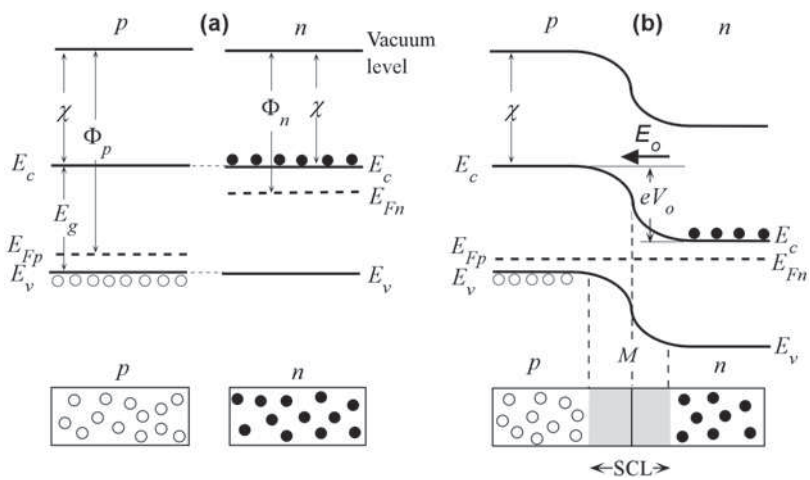


FIGURE 3.25 (a) Consider *p*- and *n*-type semiconductor (same material) before the formation of the *pn* junction, separated from each other and not interacting. (b) After the formation of the *pn* junction, there is a built-in voltage across the junction.

material properties, and obviously the same for both semiconductors. The Fermi level E_{Fp} on the *p*-side is close to E_v, and E_{Fn} on the *n*-side is close to E_c. The Fermi level is not continuous since the semiconductors are isolated. The work functions Φ_p and Φ_n are also shown.

When we bring the *p*-side and the *n*-side in contact to form the *pn* junction, we form one material system. Equilibrium in the dark in this one-material system requires that the Fermi level be uniform through the two sides of the junction, as illustrated in Figure 3.25 (b). Any change in the Fermi level across the device is equivalent to electrical work done, which must be zero in an open circuit inasmuch as the device is not generating any work and neither is there a voltage applied across it. Consider what happens when we bring the two sides together to form a *pn* junction. Far away from the metallurgical junction *M*, in the bulk of the *n*-type semiconductor we should still have an *n*-type semiconductor, and $E_c - E_{Fn}$ should be the same as in the isolated *n*-type material. Similarly, $E_{Fp} - E_v$ far away from *M* inside the *p*-type material should also be the same as in the isolated *p*-type material. These features are sketched in Figure 3.25 (b) by keeping E_{Fp} and E_{Fn} the same through the whole system and, of course, keeping the bandgap, $E_c - E_v$, the same. Clearly, to draw the energy band diagram we have to *bend the bands*, E_c and E_v, near the junction at *M* because E_c on the *n*-side is close to E_{Fn}, whereas on the *p*-side it is far away from E_{Fp}.

The instant the two semiconductors are brought together to form the junction, electrons diffuse from the *n*-side to the *p*-side and as they do so they deplete the *n*-side near the junction. Thus, E_c must move away from E_{Fn} as we move toward *M* which is exactly what is sketched in Figure 3.25 (b). Holes diffuse from the *p*-side to the *n*-side and the loss of holes in the *p*-type material near the junction means that E_v moves away from E_{Fp} as we move toward *M as* in Figure 3.25 (b). Furthermore, as electrons and holes diffuse toward each other most of them recombine and disappear around *M*, which leads to the formation of a **space charge layer** as we saw in Figure 3.15 (b). The SCL zone around the metallurgical junction has therefore been *depleted* of carriers compared with the bulk. The bending of bands E_c and E_v around *M* accounts for this depletion.

An electron in the *n*-side at E_c must overcome a potential energy (PE) barrier to go over to E_c in the *p*-side. This PE barrier is eV_o where V_o is the built-in potential, the maximum extent E_c has been bent to line up the Fermi levels. Band bending around *M* therefore accounts not only for the variation of electron and hole concentrations in this region but also for the effect of the built-in potential (and hence the built-in field, as the two are related). The diffusion of CB electrons from *n*-side to *p*-side is prevented by the built-in PE barrier eV_o. This barrier also prevents holes from diffusing from the *p*- to the *n*-side. The bending in E_c in the SCL in Figure 3.25 (b) represents the changes in the PE of the electron through this region; and hence it represents the electric field in this region. In fact $|dE_c/dx|$ would be the magnitude of the external force eE on the electron in the SCL. Note that, in the SCL region, the Fermi level is neither close to E_c nor E_v, compared with the bulk or neutral semiconductor regions. This means that both *n* and *p* in this zone are much less than their bulk values n_{no} and p_{po}. The metallurgical junction zone has been depleted of carriers compared with the bulk. Any applied voltage must therefore drop across the SCL.

The extent of band bending depends on how much we need to line up E_{Fp} and E_{Fn}, that is, $\Phi_p - \Phi_n$, as apparent when we compare Figure 3.25 (a) with (b). ($\Phi_p - \Phi_n$ is the $E_{Fn} - E_{Fp}$ difference before the contact.) The built-in potential energy eV_o is therefore $\Phi_p - \Phi_n$. The careful reader will notice that the vacuum level is no longer uniform through the whole device. This is indeed the case and represents the fact that the removal of an electron from the *n*-side will also need some additional work against the fringing electric field. (The exact analysis is quite complicated.)

B. Forward and Reverse Bias

Figure 3.26 (a) shows the energy band diagram of a *pn* junction in open circuit based on the above description (Figure 3.25). We ignored the vacuum level as this is not needed in the following discussion. When the *pn* junction is forward biased, the majority of the applied voltage drops across the depletion region so that the applied voltage is in opposition to the built-in potential, V_o. Figure 3.26 (b) shows the effect of forward bias which is to reduce the PE barrier from eV_o to $e(V_o - V)$. The electrons at E_c in the *n*-side can now readily overcome the PE barrier and diffuse to the *p*-side. The diffusing electrons from the *n*-side can be easily replenished by the negative terminal of the battery connected to this side. Similarly holes can now diffuse from the *p*- to *n*-side.

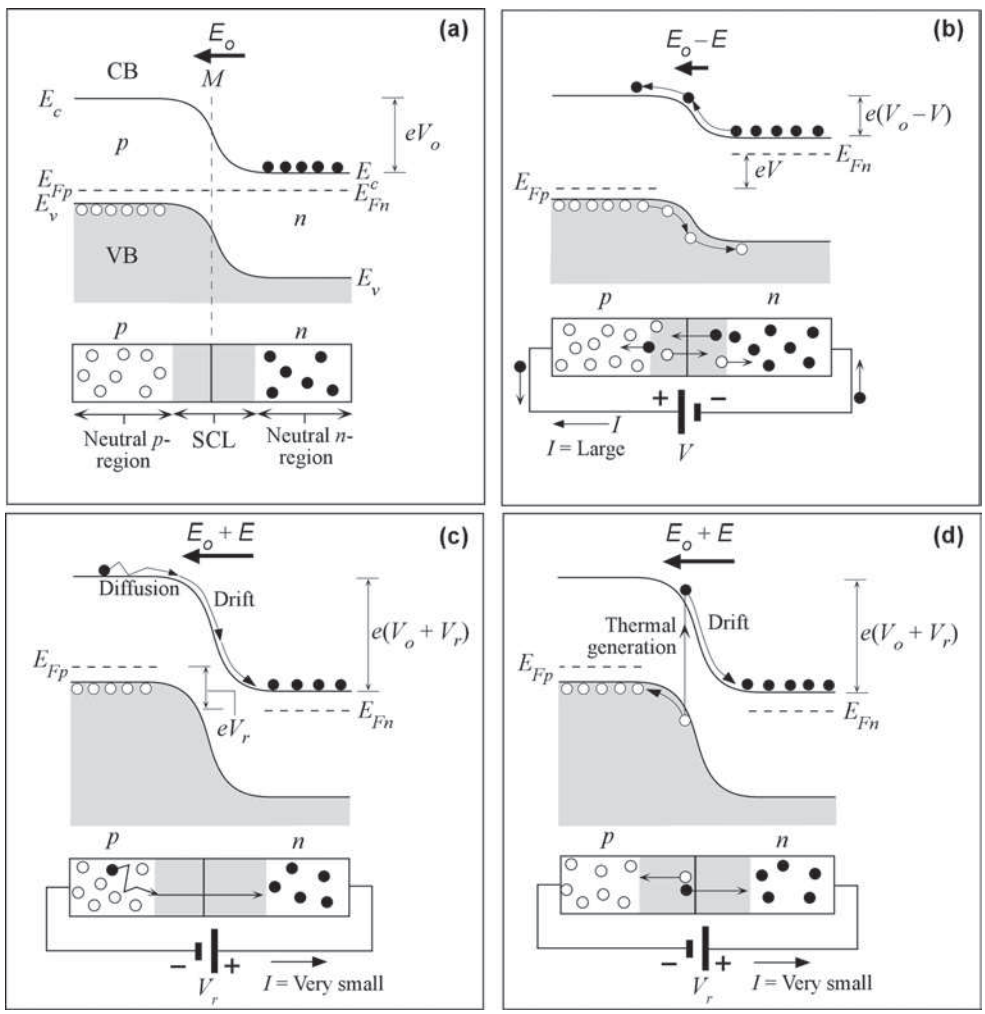


FIGURE 3.26 Energy band diagrams for a *pn* junction under (a) open circuit, (b) forward bias, and (c) and (d) reverse bias conditions. The reverse current in (c) is very small, usually negligible, since it depends on the diffusion of minority carriers to the depletion region. (d) Thermal generation of electron–hole pairs in the depletion region results in a small reverse current that is usually greater than the reverse current in (c).

The positive terminal of the battery can replenish those holes diffusing from the *p*- to the *n*-side. Therefore, a current flows through the junction and around the circuit.

The probability that an electron at E_c in the *n*-side overcomes the new PE barrier and diffuses to E_c in the *p*-side is now proportional to the Boltzmann factor $\exp\left[-e(V_o - V)/k_BT\right]$. The latter increases enormously even for small forward voltages. Thus, there is a net diffusion of electrons from the *n*- to *p*-side. Similar ideas also apply to holes at E_v in the *p*-side which also overcome the barrier $e(V_o - V)$ to diffuse into the *n*-side. Since the forward current is due to the number of electrons and holes overcoming the barrier, it is also proportional to $\exp\left[-e(V_o - V)/k_BT\right]$ or $\exp(eV/k_BT)$.

When a reverse bias, $V = -V_r$, is applied to the *pn* junction, the voltage again drops across the SCL. In this case, however, V_r adds to the built-in potential V_o so that the PE barrier becomes $e(V_o + V_r)$, as shown in Figure 3.26 (c). The field in the SCL at *M* increases to $E_o + E$ where E is the applied field (it is not simply V/W). The Shockley model predicts a small reverse saturation current due to the diffusion of minority carriers in neutral regions to the depletion region. Once they reach the depletion region, they will be drifted across by the large field and then collected by the battery. In the energy band diagram in Figure 3.26 (c), an electron in the *p*-side, within a diffusion length (L_e) to the depletion region, can diffuse to the SCL and then *fall down* the PE hill along E_c, over to the *n*-side where it will be collected by the battery. The process of falling down a PE hill is the same process as being driven by a field, in this case by $E_o + E$. The electron lost from the *p*-side is replenished from the negative terminal of the battery (or by thermal generation) to maintain the small reverse current. The same idea applies to the hole in the *n*-side within a diffusion length (L_h) to the SCL. Such minority carrier diffusion is the essence of the Shockley model. The reverse current magnitude is I_{so}. The measured reverse current in Ge diodes at room temperature follow the Shockley model as can be seen in Figure 3.22 (b).

In the case of many semiconductors, however, there is a more significant contribution to the reverse current arising from the thermal generation of electron–hole pairs in the SCL, as shown in Figure 3.26 (d), where the field here separates the pair. The electron falls down the PE hill, down E_c, to the *n*-side to be collected by the battery. Similarly the hole falls down its own PE hill (energy increases downwards for holes) to make it to the *p*-side.

EXAMPLE 3.9.1 The built-in voltage from the band diagram

Derive the expression for the built-in voltage V_o using the energy band diagram in Figure 3.25.

Solution

The extent of band bending, going from (a) to (b) in Figure 3.25, gives the PE barrier eV_o, thus

$$eV_o = \Phi_p - \Phi_n = E_{Fn} - E_{Fp} \text{ (before contact)}$$

Before the contact, on the *n*-side we have

$$n = N_c \exp\left[-(E_c - E_{Fn})/k_BT\right] = N_d$$

so that

$$E_c - E_{Fn} = -k_BT \ln(N_d/N_c) \tag{3.9.1}$$

On the *p*-side

$$n = N_c \exp\left[-(E_c - E_{Fp})/k_BT\right] = n_i^2/N_a$$

so that

$$E_c - E_{Fp} = -k_B T \ln\left[n_i^2/(N_a N_c)\right] \tag{3.9.2}$$

Thus, subtracting Eq. (3.9.2) from (3.9.1) gives

Built-in potential V_o

$$eV_o = E_{Fn} - E_{Fp} = k_B T \ln\left[(N_a N_d)/n_i^2\right] \tag{3.9.3}$$

3.10 HETEROJUNCTIONS

The *pn* junction with the band diagram in Figure 3.25 is a junction within the same crystal (Si), and hence the bandgap does not change along the device; it represents a **homojunction**. A **heterojunction** is a junction between two different semiconductor crystals with different bandgaps E_{g1} and E_{g2}, for example, between GaAs and Al_xGa_{1-x} As ternary alloys. Inasmuch as the bandgaps are now different, their alignment becomes important. If the bandgap difference is $\Delta E_g = E_{g2} - E_{g1}$, then this difference is taken up by a difference $\Delta E_c (= E_{c2} - E_{c1})$ in the CB edges, and $\Delta E_v (= E_{v1} - E_{v2})$ in the VB edges. The energy discontinuities in ΔE_c and ΔE_v are called **band offsets** and play an important role in heterojunction devices.

The terms *heterojunction* and *heterostructure* are frequently used interchangeably, though a heterostructure usually has more than one heterojunction. There may or may not be a change in the doping across the heterojucntion. The doping in the wider bandgap semiconductor is usually denoted with a capital letter *N* or *P*, and that in the narrower bandgap semiconductor with lower case *n* or *p*. There are two cases of particular importance, called Type I and Type II heterojunctions. In a Type I **straddled bandgap alignment heterojunction**, as illustrated in Figure 3.27 (a), the smaller bandgap material offers the lowest energy for both electrons and holes as in GaAs/Al_xGa_{1-x}As heterostructures in which GaAs has a smaller bandgap than Al_xGa_{1-x}As. Tpe I is the most common heterostructure in optoelectronic devices, for example, $Ga_xIn_{1-x}As/InP$, $GaAs/Ga_xIn_{1-x}P$. In a Type II **staggered lineup heterojunction**, as illustrated

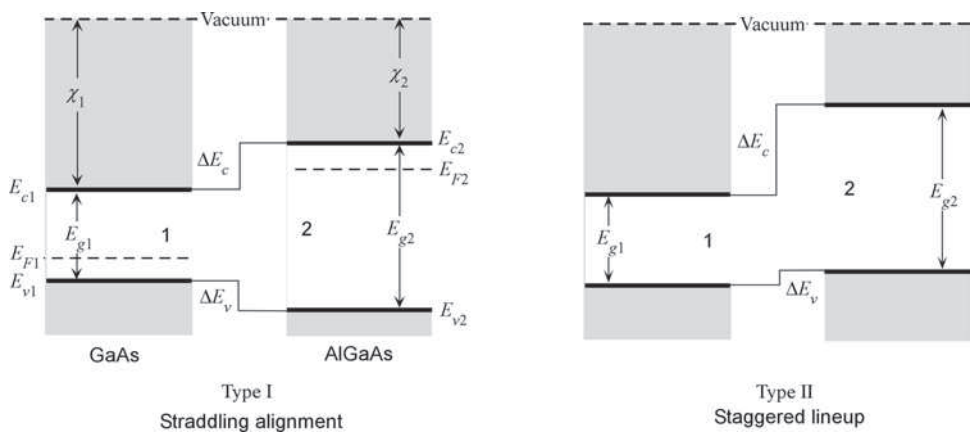


FIGURE 3.27 Two types of heterojunction and the definitions of band offsets Type I and Type II between two semiconductor crystals 1 and 2. Crystal 1 has a narrower bandgap E_{g1} than E_{g2} for crystal 2. Note that the semiconductors are not in contact so that the Fermi level in each is different. In the Type I example, crystal 1 (GaAs) is *p*-type and crystal 2 (AlGaAs) is *N*-type. (Note that the subscripts 1 and 2 refer to semiconductors on the left and right respectively.)

in Figure 3.27 (b), minimum energies for holes and electrons are in the different materials. $Ga_xIn_{1-x}As/GaAs_ySb_{1-y}$ heterojunctions over wide compositions follow the Type II behavior.

The band edge profiles for a given heterostructure are determined by the doping levels, carrier transport, and recombination around the junction, as they are for a simple *pn*-homojunction. The open-circuit heterojunction band diagrams are straightforward based on the principle that the Fermi level must be uniform once a contact is made, and we know how ΔE_g is shared between ΔE_c and ΔE_v. Figure 3.28 (a) shows the energy band diagram of an *Np* heterojunction between *n*-type AlGaAs and *p*-type GaAs. First, notice that E_F is uniform through the device as an equilibrium requirement. Far away from the junction on the *N*-side, we have an *n*-type wide bandgap AlGaAs with E_F close to E_c. Far away from the junction on the right, we have a *p*-type narrower bandgap GaAs with E_F close to E_v. In the depletion regions, around the junction, E_c and E_v must bend because there is an internal field, as we know from the ordinary *pn* junction. E_{c1} and E_{v1} of AlGaAs bend upwards and E_{c2} and E_{v2} of GaAs bend downwards as in the normal *pn* junction. We need to join E_{v1} to E_{v2} but also need to account for ΔE_v, which is easily done by putting ΔE_v between E_{v1} and E_{v2} at the junction as shown in Figure 3.28 (a). Similarly, we need to join E_{c1} to E_{c2} but also need to account for ΔE_c. In this case, we can only join E_{c1} and E_{c2} by having a narrow "spike" whose height must be ΔE_c at the junction as shown in Figure 3.28 (a).

One obvious conclusion is that the potential barrier for hole injection, $(E_{v2} - E_{v1})$, from *p* to *N* is greater than the barrier $(E_{c2} - E_{c1})$ for electron injection from *N* to *p*. Under forward bias, the current will be dominated by the injection of electrons from the *N*-side into the *p*-side. ΔE_v has increased the potential barrier against holes and ΔE_c has decreased the potential barrier against electrons.

The energy band diagram for a *pP*-type heterojunction is shown in Figure 3.28 (b). The basic principle for drawing the diagram is the same as before. In this case, there is a small spike

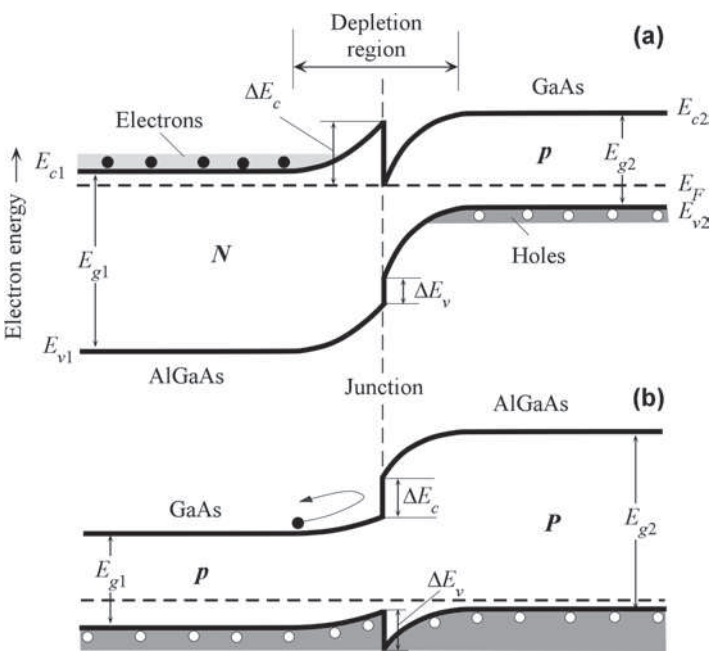


FIGURE 3.28 (a) *nP* and (b) *pP* heterojunctions and their energy band diagrams (schematic only to illustrate general features). Under open circuit and equilibrium conditions, the Fermi level E_F must be uniform, that is, continuous throughout the device. If E_F is close to the conduction band edge, E_c, it results in an *n*-type, and if it is close to the valence band edge, E_v, it results in a *p*-type semiconductor. There is a discontinuity ΔE_c as in E_c, and ΔE_v in E_v, right at the junction.

in E_v at the junction. ΔE_c increases the potential barrier from E_{c1} to E_{c2}. An electron in the CB in the p-side cannot simply overcome this barrier and enter the P-side. Holes in the VB of p- and P-sides can easily cross through the spike by tunneling. Both the Np and pP heterojunctions are used extensively in LED and semiconductor laser diode heterostructures.

3.11 LIGHT-EMITTING DIODES: PRINCIPLES

A. Homojunction LEDs

A light-emitting diode is essentially a pn junction diode typically made from a direct bandgap semiconductor, for example, GaAs, in which the electron–hole pair recombination results in the emission of a photon. The emitted photon energy is therefore approximately equal to the bandgap energy, $h\upsilon \approx E_g$. Figure 3.29 (a) shows the energy band diagram of an unbiased pn^+ junction device in which the n-side is more heavily doped than the p-side. The band diagram is drawn to keep the Fermi level, E_{Fp} and E_{Fn} on the p- and n-sides, uniform through the device which is a requirement of equilibrium with no applied bias. The depletion region in a pn^+ device extends mainly into the p-side. There is a potential energy (PE) barrier eV_o from E_c on the n-side to E_c on the p-side, that is, $\Delta E_c = eV_o$, where V_o is the *built-in voltage.* The higher concentration of conduction (free) electrons in the n-side encourages the diffusion of these electrons from the n- to the p-side. This net electron diffusion, however, is prevented by the electron PE barrier eV_o.

As soon as a forward bias V is applied, this voltage drops almost entirely across the depletion region since this is the most resistive part of the device. Consequently, the built-in potential V_o is reduced to $V_o - V$, which then allows the electrons from the n^+ side to diffuse, or become injected, into the p-side as illustrated in Figure 3.29 (b). The hole injection component from p into the n^+ side is much smaller than the electron injection component from the n^+ to p-side. The recombination of injected electrons in the depletion region as well as in the neutral p-side results in the *spontaneous emission* of photons. Recombination primarily occurs within the depletion region and within a volume extending over the diffusion length L_e of the electrons in the p-side. (Electron injection is preferred over hole injection in GaAs LEDs because electrons have a higher mobility and hence a larger diffusion coefficient.) This recombination zone is

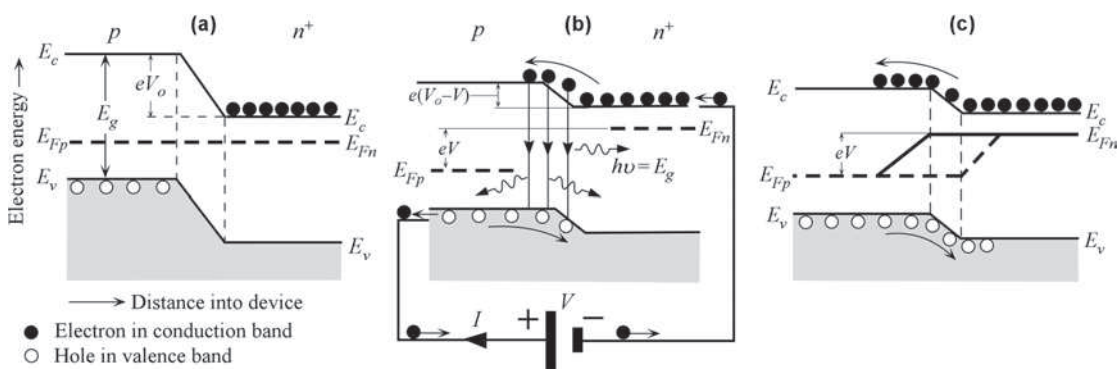


FIGURE 3.29 (a) The energy band diagram of a pn^+ (heavily n-type doped) junction without any bias. Built-in potential V_o prevents electrons from diffusing from n^+ to p-side. (b) The applied bias potential V reduces V_o and thereby allows electrons to diffuse, be injected, into the p-side. Recombination around the junction and within the diffusion length of the electrons in the p-side leads to spontaneous photon emission. (c) Quasi-Fermi levels E_{Fp} and E_{Fn} for holes and electrons across a forward-biased pn-junction.

frequently called the **active region**.[16] The phenomenon of light emission from EHP recombination as a result of minority carrier injection, as in this case, is called **injection electroluminescence**.[17] Because of the statistical nature of the natural recombination process between electrons and holes, the emitted photons are in random directions; such randomly emitted radiation is called **spontaneous emission**. The LED structure has to be such that the emitted photons can escape the device without being reabsorbed by the semiconductor material. This means the *p*-side has to be sufficiently narrow, or we have to use heterostructure devices as discussed below. Further, the rate of recombination depends on the product *np*, which must be increased to enhance the emitted photon flux. We need to confine the carriers and increase the carrier concentration, again achievable by using heterostructures.

Notice that in Figure 3.29 (b), under forward bias, E_{Fn} and E_{Fp} are separated by eV inasmuch as the electrical work done per electron, eV, must be ΔE_F or $E_{Fn} - E_{Fp}$. Also notice that E_{Fn} and E_{Fp} extend into the depletion region, where there are now two Fermi levels as highlighted in Figure 3.29 (c). We can use E_{Fn} to represent the electron concentration through the general expression $n = N_c \exp\left[-(E_c - E_{Fn})/k_B T\right]$, not only on the *n*-side but also in the depletion region, and even in the *p*-side; E_{Fn} is then called a **quasi-Fermi level for electrons**. Similarly, we can use E_{Fp} as a **quasi-Fermi level for holes** and use $p = N_v \exp\left[-(E_{Fp} - E_v)/k_B T\right]$ to represent the hole concentration.[18] In the depletion region, E_{Fn} and E_{Fp} are separated within the same spatial region so that $np > n_i^2$, which is expected given that we are injecting minority carriers and have non-equilibrium. Figure 3.29 (c) shows the variation of E_{Fp} and E_{Fn} across a biased *pn* junction. Such variation in E_{Fp} and E_{Fn} will result in $p(x)$ and $n(x)$ profiles across the device that match the sketch in Figure 3.16 (a) for a pn^+ junction; we have effectively represented carrier concentration profiles in terms of E_{Fp} and E_{Fn} on the energy band diagram. Notice that E_{Fn} on the *p*-side slopes down, away from E_c, to represent the decrease in the electron concentration in the neutral *p*-side, due to recombination of injected electrons with the majority carriers (holes). E_{Fn} reaches E_{Fp} over a distance that is very roughly the diffusion length of injected electrons. The region where E_{Fn} and E_{Fp} have merged and are the same represents the neutral bulk *p*-side. Similar arguments apply to holes injected into the *n*-side. E_{Fp} slopes and merges with E_{Fn} over a distance that is roughly the hole diffusion length (shorter than the electron diffusion length).

It is clear that the emission of photons in the LED involves the direct recombination of electrons and holes; that is, an electron makes a radiative transition from the bottom of the CB to the top of the VB, where there is an empty state (a hole) as in Figure 3.14 (a). It may therefore be thought that one should avoid using indirect semiconductors in LEDs. However, this is not entirely true because it is possible to introduce impurities into certain indirect semiconductors and thereby enable radiative transitions through these impurities. A good example is GaP doped with N, that is, GaP:N, which are used in inexpensive green LEDs in numerous indicator and display applications in consumer optoelectronics. GaP is an indirect bandgap semiconductor with E_g of 2.26 eV (corresponding to green light at 550 nm). When N is added, N substitutes for P,

[16] The "active region" term is probably more appropriate for laser diodes in which there is photon amplification in this region. However, the term is also used in LED discussions.

[17] In 1907, Henry Round, working for Guglielmo Marconi in England, observed yellow light emission when a current was passed through a metal-SiC (silicon carbide or corborandum) crystal rectifier. Oleg Losev in the former USSR carried out a number of experiments on the emission of light upon passing a current through such metal-SiC rectifiers during the mid-1970s. He reported his systematic experiments in a number of journals, called the effect the "inverse photoelectric effect," and proposed that emission frequency is $\upsilon = eV/h$. (See Nikolay Zheludev, *Nat. Photonics*, *1*, 189, 2007.)

[18] Clearly, quasi-Fermi levels provide a convenient and useful means of representing carrier concentrations on the energy diagram.

and is called an **isoelectronic dopant**. However, N dopants create recombination centers that have a localized energy level at E_r, roughly $0.1 - 0.2$ eV below the CB as shown in Figure 3.14 (c). An electron at the bottom of the CB can easily fall into E_r during which phonons are emitted to make up for the change in the momentum from k_{CB} to k_{VB}. The electron at E_r can then drop down to the top of the VB and emit a photon $h\upsilon = E_r - E_v$, with $h\upsilon$ slightly below E_g. Since $h\upsilon$ is less than E_g, there is less likelihood of reabsorbing the photon in the bulk of GaP, especially if N doping is introduced in the junction (recombination) region only.

B. Heterostructure High Intensity LEDs

The *pn*-junction LED shown in Figure 3.29 suffers from a number of drawbacks and has a low efficiency. The *p*-region must be narrow to allow the photons to escape without much reabsorption. When the *p*-side is narrow, some of the injected electrons in the *p*-side reach the surface by diffusion and recombine through crystal defects near the surface. This radiationless recombination process decreases the light output. In addition, if the recombination occurs over a relatively large volume (or distance), due to long electron diffusion lengths, then the chances of reabsorption of emitted photons becomes higher; the amount of reabsorption increases with the material volume.

On the other hand, heterostructure LEDs can have exceptionally high efficiencies along with a number of other advantages; most modern LEDs are heterostructure devices (HDs). First, we note that the refractive index of a semiconductor material depends on its bandgap. A wider bandgap semiconductor has a lower refractive index. This means that by constructing LEDs from heterostructures, we can engineer a dielectric waveguide within the device and thereby channel photons out from the recombination region.

LED constructions for increasing the intensity of the output light make use of the double heterostructure (DH) structure. Figure 3.30 (a) shows a **double-heterostructure** (DH) device based on two heterojunctions between different semiconductor crystals with different bandgaps. In this case the semiconductors are AlGaAs ($Al_xGa_{1-x}As$) with $E_g \approx 2$ eV and GaAs with $E_g \approx 1.4$ eV. The double heterostructure in Figure 3.30 (a) has an N^+p heterojunction between N^+-AlGaAs and *p*-GaAs. There is another heterojunction between *p*-GaAs and *P*-AlGaAs. The *p*-GaAs region is a thin layer, typically a fraction of a micron and it is lightly doped. The reader would have noticed that this HD device is made of the heterojunctions in Figure 3.28 (a) and (b) one after the other to form an *N-p-P* device.

The simplified energy band diagram for the whole device in the absence of an applied voltage is shown in Figure 3.30 (b). The Fermi-level E_F is continuous through the whole structure. There is a potential energy barrier eV_o for electrons in the CB of N^+-AlGaAs against diffusion into *p*-GaAs. The spike in E_c at the junction, as in Figure 3.28 (a), is ignored in this simplified sketch. There is a bandgap change at the junction between *p*-GaAs and *P*-AlGaAs which results in a step change ΔE_c in E_c between the two CBs of *p*-GaAs and *P*-AlGaAs as in Figure 3.28 (b). This step change, ΔE_c, is effectively a potential energy barrier that prevents any electron in the CB in *p*-GaAs passing into the CB of *P*-AlGaAs.

When a forward bias is applied, the majority of this voltage drops between the N^+-AlGaAs and *p*-GaAs (*i.e.,* across the depletion region) and reduces the potential energy barrier eV_o, just as in the normal *pn* junction diode. This allows electrons in the CB of N^+-AlGaAs to be injected (by diffusion) into *p*-GaAs as shown in Figure 3.30 (c). These electrons, however, are *confined* to the CB of *p*-GaAs since there is a barrier ΔE_c between *p*-GaAs and *P*-AlGaAs (see Figure 3.28 (b)). The wide bandgap AlGaAs layer therefore acts as a **confining layer** that restricts injected electrons to the *p*-GaAs layer. The recombination of injected electrons with the holes present in

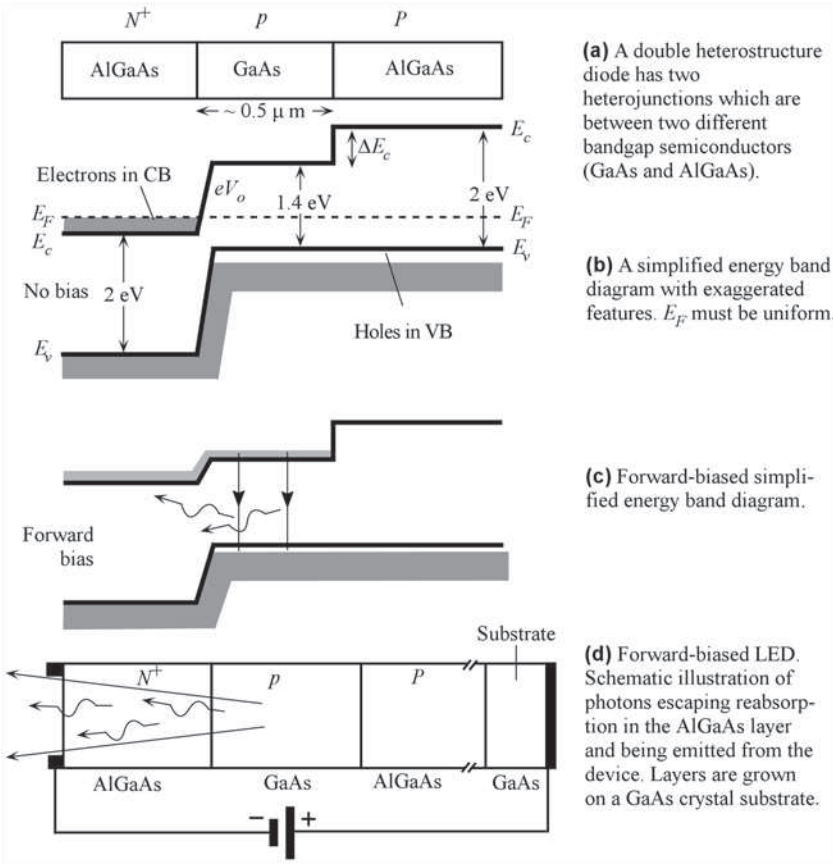
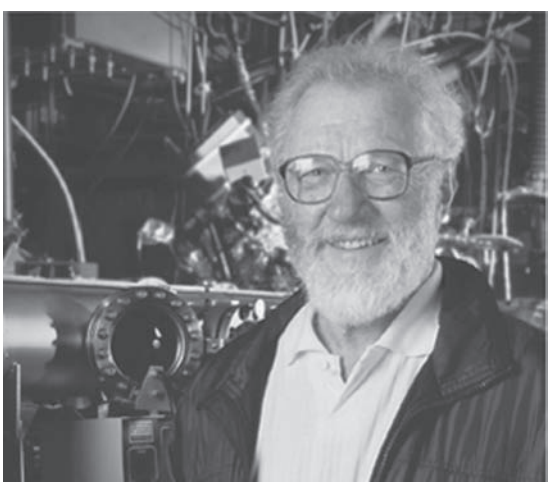


FIGURE 3.30 The double heterostructure AlGaAs/GaAs/AlGaAs LED structure (highly simplified schematic).



Herbert Kroemer (left), along with Zhores Alferov (Chapter 4 opening photo), played a key role in the development of semiconductor heterostructures that are widely used in modern optoelectronics.[19] Herbert Kroemer was also well-recognized for his experimental work on the fabrication of heterostructures by using an atomic layer-by-layer crystal growth technique called Molecular Beam Epitaxy (MBE)—the equipment shown behind Professor Kroemer in the photo. Since 1976, Professor Kroemer has been with the University of California, Santa Barbara, where he continues his research. Herbert Kroemer and Zhores Alferov shared the Nobel Prize in Physics (2000) with Jack Kilby. Their Nobel citation is "*for developing semiconductor heterostructures used in high-speed- and opto-electronics.*" *(Courtesy of Herbert Kroemer, University of California, Santa Barbara.)*

[19] See H. Kroemer, *Rev. Mod. Phys.*, *73*, 783, 2001.

this p-GaAs layer results in spontaneous photon emission. Since the bandgap E_g of AlGaAs is greater than that of GaAs, the emitted photons do not get reabsorbed as they escape the active region and can reach the surface of the device as illustrated in Figure 3.30 (d). Since light is also not absorbed in P-AlGaAs, it can be *reflected* to increase the light output, by, for example, using a dielectric mirror at the end of P-AlGaAs. Another advantage of the AlGaAs/GaAs heterojunction is that there is only a small lattice mismatch between the two crystal structures and hence negligible strain induced interfacial defects (*e.g.*, dislocations) in the device compared with the defects at the surface of the semiconductor in conventional homojunction LED structure. The DH LED is much more efficient than the homojunction LED.

C. Output Spectrum

The energy of an emitted photon from an LED is not simply equal to the bandgap energy E_g because electrons in the conduction band are distributed in energy and so are the holes in the valence band. Suppose that the active region is p-type, and excess electrons have been injected by forward bias. Figure 3.31 (a) and (b) illustrates the energy band diagram and the energy distributions of electrons and holes in the CB and VB, respectively, for a p-type semiconductor. The electron concentration as a function of energy in the CB is given by $g(E)f(E)$, where $g(E)$ is the density of states in the CB and $f(E)$ is the Fermi–Dirac function (probability of finding an electron in a state with energy E). The product $g(E)f(E)$ represents the electron concentration per unit energy, $n_E(E)$, and is plotted along the horizontal axis in Figure 3.31 (b). There is a similar energy distribution for holes, p_E, in the VB but p_E is enormously larger than n_E. The E–k diagram for a typical direct bandgap semiconductor (such as GaAs) is shown in Figure 3.31 (c). Since the hole concentration is very large, we can assume that the rate of recombination will depend primarily on the concentration of injected electrons.[20] The electron concentration in the CB

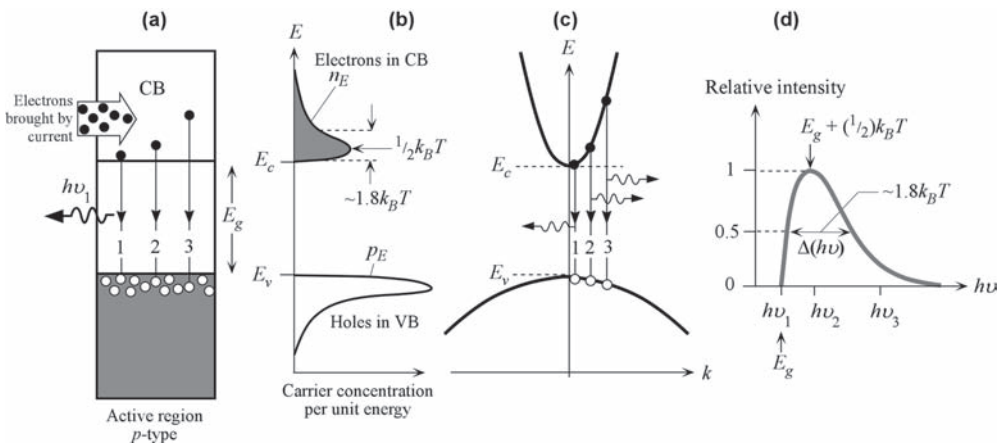


FIGURE 3.31 (a) Energy band diagram with possible recombination paths. (b) Energy distribution of electrons in the CB and holes in the VB. The highest electron concentration is $(1/2)k_BT$ above E_c. (c) A simplified E–k (equivalent to energy vs. momentum) diagram and direct recombination paths in which k (*i.e.,* momentum) is conserved. (d) The relative light intensity as a function of photon energy based on (b) and (c).

[20] We also need to consider the quantum mechanical transition probability from an occupied state in the CB to an empty state in the VB but, for simplicity, we will take this to be constant, that is, we will ignore it.

as a function of energy is asymmetrical, and has a peak at $\frac{1}{2}k_BT$ above E_c. The energy spread of these electrons is about $1.8k_BT$ between the half-maximum points, as in Figure 3.31 (b), or roughly $\sim 2k_BT$. When an electron at E_c recombines with a hole at E_v, shown as the transition 1 in Figure 3.31 (a), a photon is emitted with an energy $h\upsilon_1 = E_c - E_v = E_g$. Since there are not many electrons and holes at the band edges, this type of recombination does not occur frequently, and the emitted light intensity from a type 1 transition is small.

The transition that involves the largest electron concentration (the peak in n_E) is shown as 2 in Figure 3.31 (a), and emits a photon with $h\upsilon_2 > h\upsilon_1$. Such transitions occur frequently (large n_E), and hence the emitted intensity from type 2 transitions is much larger than that for type 1. Similarly, the transition marked 3, corresponding to $h\upsilon_3 > h\upsilon_2$, involves an electron quite high up in the CB where n_E is very small. Such type 3 transitions are infrequent and lead to a small emission intensity. The emission intensity therefore rises to a maximum and then falls with $h\upsilon$ as shown in Figure 3.31 (d).

One might conclude that the highest emitted intensity should intuitively correspond to the transition from the peak in n_E to the peak in p_E in Figure 3.31 (b) that emits $h\upsilon = E_g + k_BT$. However, we also need to consider the conservation of momentum, and hence the E–k diagram in Figure 3.31 (c). The momentum of the emitted photon is negligibly small, so that an electron falls straight down in the E–k diagram without changing its k-vector, that is, the electron momentum $\hbar k$ is conserved. The E–k curvatures are different in the CB and the VB. The electron at $E_c + \frac{1}{2}k_BT$ cannot just recombine with the hole at $E_v + \frac{1}{2}k_BT$ because that transition does not satisfy the $\hbar k$-conservation. As shown in Figure 3.31 (c), direct recombination involves energetic electrons spreading over several k_BT in the CB, more than the holes in the VB, because the E–k curvature is narrower in the CB and broader in the VB. It is apparent that the emission spectrum in this case is determined by n_E, the energy spread in the electrons in the CB, so that the emission has a peak at roughly $E_g + \frac{1}{2}k_BT$. Further the spread $\Delta(h\upsilon)$ in the emitted photon energies should be roughly the spread in n_E, that is, $\Delta(h\upsilon) \approx 1.8k_BT$.

The intuitive relative light intensity vs. photon energy characteristic of the output spectrum based on n_E and the E–k diagram is shown in Figure 3.31 (d) and represents an important LED characteristic. Given the spectrum in Figure 3.31 (d), we can also obtain the relative light intensity vs. wavelength characteristic since $\lambda = c/\upsilon$, which would look like Figure 3.31 (d) flipped horizontally. The **linewidth** of the output spectrum, $\Delta\upsilon$ or $\Delta\lambda$, is defined as the width between half-intensity points as illustrated in Figure 3.31 (d).

The experimentally observed output spectrum, that is, the relative intensity vs. wavelength characteristics, from an LED depends not only on the semiconductor material, including dopant concentrations, but also on the structure of the *pn* junction diode. The spectrum in Figure 3.31 (d) represents a highly idealized spectrum without including the effects of heavy doping on the energy bands. For a heavily doped *n*-type semiconductor there are so many donors that the electron wavefunctions at these donors overlap to generate a narrow impurity band centered at E_d but extending into the conduction band. Thus, the donor impurity band overlaps the conduction band and hence effectively lowers E_c as in Figure 3.10 (a). The minimum emitted photon energy from heavily doped semiconductors is therefore less than E_g and depends on the amount of doping.

Typical output spectrum from an AlGaAs IR LED is shown in Figure 3.32 (a). Notice that the spectrum exhibits significantly less asymmetry than the idealized spectrum in Figure 3.31 (d). The width of the spectrum is about 40 nm, which corresponds to a width of about $2.9k_BT$ in the energy distribution of the emitted photons, more than the expected $1.8k_BT$. The reasons for not observing the theoretical spectrum in Figure 3.31 (d) are essentially twofold. First, higher energy photons become reabsorbed in the material and photogenerate electrons and holes.

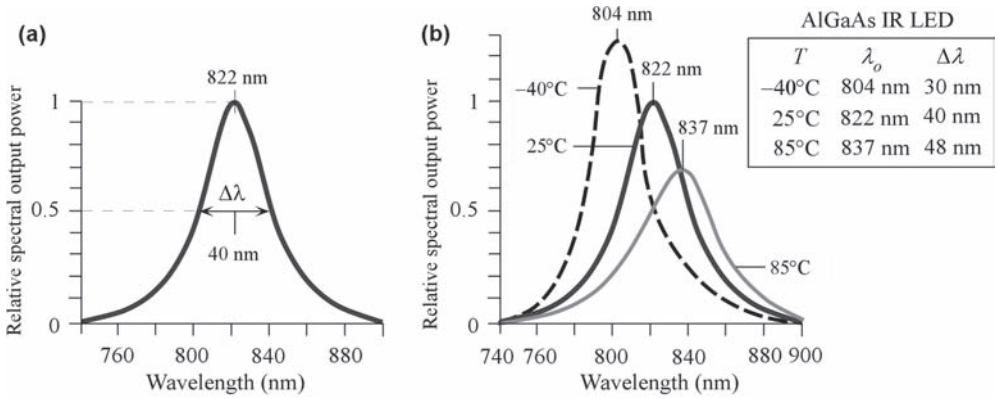


FIGURE 3.32 (a) A typical output spectrum (relative intensity vs. wavelength) from an IR (infrared) AlGaAs LED. (b) The output spectrum of the LED in (a) at three temperatures: 25°C, −40°C, and 85°C. Values normalized to peak emission at 25°C. The spectral widths are FWHM.

These electrons and holes thermalize and end up recombining to emit photons with lower energies, closer to E_g. Thus photons become redistributed. Secondly, the band edges E_c and E_v are not sharp in heavily doped semiconductors, which leads to the smearing of the well-defined E_g for the emission onset. The peak emission frequency υ_o and the spectral width $\Delta\upsilon$ in photon energy in LEDs with a direct bandgap active region are normally described by

LED spectrum in frequency

$$h\upsilon_o \approx E_g + \tfrac{1}{2}k_BT \quad \text{and} \quad h\Delta\upsilon = mk_BT \tag{3.11.1}$$

where m is a numerical factor that is typically between 1.5 and 3.5, though for many LEDs, $m \approx 3$ is a good value (see Question 3.23). The corresponding peak wavelength λ_o and the spectral width $\Delta\lambda$ can be easily found from Eq. (3.11.1) as in Example 3.11.1. Further, the actual position of the peak is likely to be somewhat more than $\frac{1}{2}k_BT$ in Eq. (3.11.1), given a broader observed spread than $1.8k_BT$ expected from the theory. As the temperature increases, the change in $h\upsilon_o$ in Eq. (3.11.1) is due mainly to the decrease in the bandgap E_g with temperature. The peak emission wavelength λ_o, corresponding to υ_o, therefore increases with temperature as shown Figure 3.32 (b). In addition, the linewidth $\Delta\lambda$ becomes longer at higher temperatures as electrons are distributed further into the CB. Thus, a wider spectrum of photon energies are emitted upon EHP recombination. The variation of the bandgap E_g with temperature is commonly represented by the **Varshni equation**

Varshni bandgap equation

$$E_g = E_{go} - \frac{AT^2}{B + T} \tag{3.11.2}$$

where E_{go} is the bandgap at $T = 0$ K, and A and B are material-specific constants that are listed in various semiconductor handbooks.

Equation (3.11.1) does not apply to an indirect bandgap semiconductor in which a recombination center is involved in the radiative transition, such as GaP:N. The electron localized at the recombination center would have a significant uncertainty in its momentum Δp and hence an uncertainty ΔE in its energy (Heisenberg's uncertainty principle, $\Delta p\Delta x \sim \hbar$). The emitted photon spectrum depends on this ΔE, and is wider than $3k_BT$ that is involved in the direct recombination process in Figure 3.31.

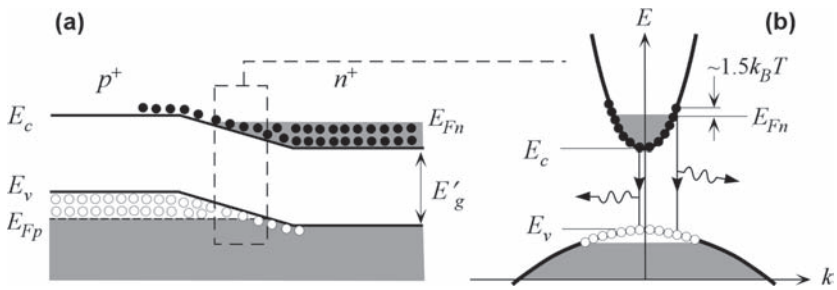


FIGURE 3.33 (a) Forward-biased degenerately doped *pn* junction. The bandgap E_g' is narrower than that in the undoped bulk crystal. The quasi-Fermi levels E_{Fn} and E_{Fp} overlap around the junction. (b) The transitions involved in a degenerately doped *pn* junction.

For degenerately doped junctions, the Fermi level E_{Fn} on the *n*-side and E_{Fp} on the *p*-side will be in the CB and VB, respectively. Under a large forward bias, the active region can have E_{Fn} in the CB, E_{Fp} in the VB around the junction as shown in Figure 3.33 (a). The bandgap E_g' is narrower than E_g in the undoped crystal (see Figure 3.10 (a) and (b)). As shown in the E–k diagram in Figure 3.33 (b), electrons occupy states from the CB edge E_c up to about $\sim 1.5k_BT$ above E_{Fn}. The emission spectrum will extend from $h\upsilon \approx E_g$ to about $E_g + E_{Fn} + 1.5k_BT$ so the width is roughly $E_{Fn} + 1.5k_BT$.

EXAMPLE 3.11.1 LED spectral linewidth

We know that a spread in the output wavelengths is related to a spread in the emitted photon energies as illustrated in Figure 3.31. The emitted photon energy $h\upsilon = hc/\lambda$. Assume that the spread in the photon energies $\Delta(h\upsilon) \approx 3kT$ between the half intensity points. Show that the corresponding linewidth $\Delta\lambda$ between the *half intensity points* in the output spectrum is

$$\Delta\lambda = \lambda_o^2 \frac{3kT}{hc} \tag{3.11.3}$$

LED spectral linewidth

where λ_o is the peak wavelength. What is the spectral linewidth of an optical communications LED operating at 1310 nm and at 300 K?

Solution

First consider the relationship between the photon frequency υ and λ,

$$\lambda = \frac{c}{\upsilon} = \frac{hc}{h\upsilon}$$

in which $h\upsilon$ is the photon energy. We can differentiate this

$$\frac{d\lambda}{d(h\upsilon)} = -\frac{hc}{(h\upsilon)^2} = -\frac{\lambda^2}{hc} \tag{3.11.4}$$

The negative sign implies that increasing the photon energy decreases the wavelength. We are only interested in changes or spreads, thus $\Delta\lambda/\Delta(h\upsilon) \approx |d\lambda/d(h\upsilon)|$, and this spread should be around $\lambda = \lambda_o$, so that Eq. (3.11.4) gives

$$\Delta\lambda = \frac{\lambda_o^2}{hc}\Delta(h\upsilon) = \lambda_o^2 \frac{3kT}{hc}$$

where we used $\Delta(h\nu) = 3kT$. We can substitute $\lambda = 1310$ nm and $T = 300$ K to calculate the linewidth of the 1310 nm LED

$$\Delta\lambda = \lambda^2 \frac{3kT}{hc} = (1310 \times 10^{-9})^2 \frac{3(1.38 \times 10^{-23})(300)}{(6.626 \times 10^{-34})(3 \times 10^8)} = 1.07 \times 10^{-7}\,\text{m} \quad \text{or} \quad 107\,\text{nm}$$

The spectral linewidth of an LED output is due to the spread in the photon energies, which is fundamentally about $3kT$. The only option for decreasing $\Delta\lambda$ at a given wavelength is to reduce the temperature. The output spectrum of a laser, on the other hand, has a much narrower linewidth.

EXAMPLE 3.11.2 LED spectral width

Consider the three experimental points in Figure 3.32 (b) as a function of temperature. By a suitable plot find m and verify Eq. (3.11.3).

Solution

From Example 3.11.1, we can use Eq. (3.11.3) with m instead of 3 as follows

LED linewidth and temperature

$$\frac{\Delta\lambda}{\lambda_o^2} = \left(\frac{mk}{hc}\right)T \tag{3.11.5}$$

and plot $\Delta\lambda/\lambda_o^2$ vs. T. The slope of the best line forced through zero should give mk/hc and hence m. Using the three λ_o and $\Delta\lambda$ values in the inset of Figure 3.32 (b), we obtain the graph in Figure 3.34. The best line is forced through zero to follow Eq. (3.11.5), and gives a slope of $1.95 \times 10^{-7}\,\text{nm}^{-1}\,\text{K}^{-1}$ or $195\,\text{m}^{-1}\,\text{K}^{-1}$. Thus,

$$\text{Slope} = 195\,\text{m K}^{-1} = \frac{m(1.38 \times 10^{-23}\,\text{J K}^{-1})}{(6.626 \times 10^{-34}\,\text{J s})(3 \times 10^8\,\text{m s}^{-1})}$$

so that

$$m = 2.81$$

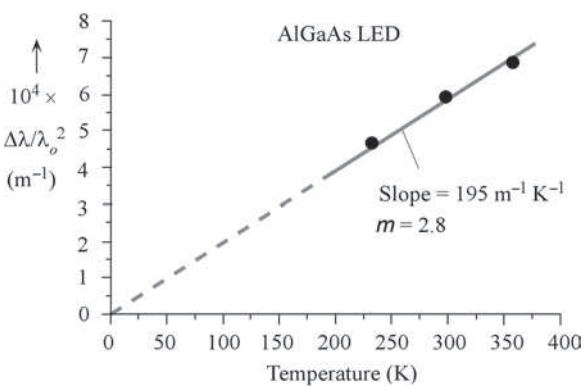
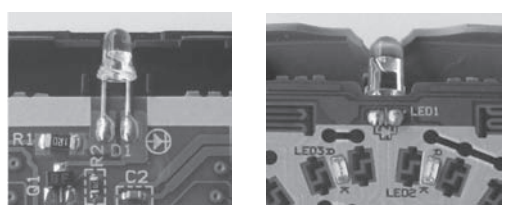


FIGURE 3.34 The plot of $\Delta\lambda/\lambda_o^2$ vs. T for an AlGaAs infrared LED, using the peak wavelength λ_o and spectral width $\Delta\lambda$ at three different temperatures, using the data shown in Figure 3.32 (b).



Infrared LEDs are widely used in various remote controls.

EXAMPLE 3.11.3 Dependence of the emission peak and linewidth on temperature

Using the Varshni equation, Eq. (3.11.2), find the shift in the peak wavelength (λ_o) emitted from a GaAs LED when it is cooled from 25°C to −25°C. The Varshni constants for GaAs are $E_{go} = 1.519\,\text{eV}$, $A = 5.41 \times 10^{-4}\,\text{eV K}^{-1}$, $B = 204\,\text{K}$.

Solution

At $T = 298\,\text{K}$, using the Varshni equation

$$E_g = E_{go} - AT^2/(B + T)$$
$$= 1.519\,\text{eV} - (5.41 \times 10^{-4}\,\text{eV K}^{-1})(298\,\text{K})^2/(204\,\text{K} + 298\,\text{K}) = 1.423\,\text{eV}$$

At 298 K, $(1/2)k_BT = 0.0128$ eV. The peak emission is at $h\upsilon_o \approx E_g + (1/2)k_BT$. Using $\upsilon_o = c/\lambda_o$, we get

$$\lambda_o = \frac{ch}{(E_g + \frac{1}{2}k_BT)} = \frac{(3 \times 10^8\,\text{m s}^{-1})(6.626 \times 10^{-34}\,\text{J s})/(1.602 \times 10^{-19}\,\text{eV J}^{-1})}{(1.4223\,\text{eV} + 0.0128\,\text{eV})} = 864.2\,\text{nm}$$

At −25°C, or 248 K, $(1/2)\,k_BT = 0.0107$ eV, repeating the above calculation

$$E_g = 1.519\,\text{eV} - (5.41 \times 10^{-4}\,\text{eV K}^{-1})(248\,\text{K})^2/(204\,\text{K} + 248\,\text{K}) = 1.445\,\text{eV}$$

and the new peak emission wavelength λ_o' is

$$\lambda_o' = \frac{(3 \times 10^8\,\text{m s}^{-1})(6.626 \times 10^{-34}\,\text{J s})/(1.602 \times 10^{-19}\,\text{eV J}^{-1})}{(1.445\,\text{eV} + 0.0107\,\text{eV})} = 852.4\,\text{nm}$$

The change $\Delta\lambda = \lambda_o - \lambda_o' = 864.2 - 852.4 = 11.8\,\text{nm}$ over 50°C, or 0.24 nm/°C. The examination of Figure 3.32 (b) shows that the change in the peak wavelength per unit temperature in the range −40°C to 85°C is roughly the same. Because of the small change, we kept four significant figures in E_g and λ_o calculations.

3.12 QUANTUM WELL HIGH INTENSITY LEDS

A typical **quantum well** (QW) device has an ultra-thin, typically less than 50 nm, narrow bandgap semiconductor with a bandgap E_{g1} sandwiched between two wider bandgap semiconductors with a bandgap E_{g2}, as illustrated in Figure 3.35 (a). For example, this could be a thin GaAs (E_{g1}) layer sandwiched between two $Al_xGa_{1-x}As$ (E_{g2}) layers. The wide bandgap layers are called **confining** layers. We assume that the two semiconductors are lattice matched in the sense that they have the same crystal structure and lattice parameter *a.* This means that interface defects due to the mismatch of crystal dimensions between the two semiconductor crystals are minimal; and neglected. Since the bandgap, E_g, changes at the interface, there are discontinuities in E_c and E_v at the interfaces as before; these discontinuities, ΔE_c and ΔE_v, are shown in Figure 3.35 (b) and depend on the semiconductor properties.[21] Because of the potential energy barrier, ΔE_c, conduction electrons in the thin E_{g1}-layer are confined in the *x*-direction. This confinement length *d*, the width of the thin E_{g1}-semiconductor, is so small that we can treat the electron as

[21] In the case of GaAs/AlGaAs heterostructure, ΔE_c is greater than ΔE_v. Very approximately, the change from the wider E_{g2} to narrower E_{g1} is proportioned 60% to ΔE_c and 40% to ΔE_v.

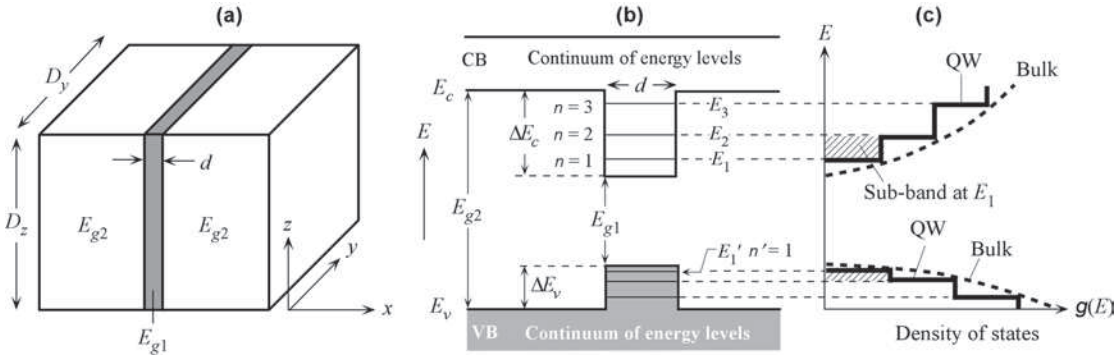


FIGURE 3.35 (a) A single quantum well (SQW) of a smaller bandgap material (E_{g1}) of thickness d along x surrounded by a thicker material of wider bandgap (E_{g2}). (b) The electron energy levels associated with motion along x are quantized as E_1, E_2, E_3, *etc.* Each level is characterized by a quantum number n. (c) The density of states for a bulk semiconductor and a QW.

in a one-dimensional (1D) potential energy (PE) well in the x-direction but as if it were free in the yz plane.

The energy of the electron in the QW must reflect its 1D quantization in the x-direction, and its freedom in the yz plane. If E_n is the electron energy in the well, then

Energy in a 1D quantum well

$$E_n = E_c + \frac{h^2 n^2}{8m_e^* d^2} + \frac{\hbar^2 k_y^2}{2m_e^*} + \frac{\hbar^2 k_z^2}{2m_e^*} \qquad \textbf{(3.12.1)}$$

where n is a quantum number having the values 1, 2, 3,…, and k_y and k_z are the wave vectors of the electron along y- and z-directions. The reason for the E_c in Eq. (3.12.1) is that the potential energy barriers are defined with respect to E_c. These PE barriers are ΔE_c along x and electron affinity (energy required to take the electron from E_c to vacuum) along y and z. The second term is the energy of an electron in an infinite PE well, whereas we have a finite PE well of depth ΔE_c. Thus, the second term is only an approximation. The minimum energy E_1 corresponds to $n = 1$ and is above E_c of the E_{g1}-semiconductor as shown in Figure 3.35 (b). For any given n value, we have a sub-band of energies due to k_y and k_z terms in Eq. (3.12.1). The separation between the energy levels associated with motion in the yz plane in a sub-band is so small that the electron is free to move in the yz plane as if it were in the bulk semiconductor; we assume a continuum of energy as shown in Figure 3.36. We therefore have a *two-dimensional electron gas* which is confined in the x-direction. The holes in the valence band are confined by the potential energy barrier ΔE_v (hole energy is in the opposite direction to electron energy) and behave similarly as illustrated in Figure 3.35 (b). They are characterized by the quantum number $n' = 1, 2$, *etc.* corresponding to the levels E_1', E_2', *etc.* [only E_1' is marked in Figure 3.35 (b)].

The density of electronic states for the two-dimensional electron system is not the same as that for the bulk semiconductor. For a given electron concentration n, the density of states $g(E)$, the number of quantum states per unit energy per unit volume, is constant and does not depend on the electron energy. The density of states for the confined electron and that in the bulk semiconductor are shown schematically in Figure 3.35 (c). $g(E)$ is constant at E_1 until E_2, where it increases as a step and remains constant until E_3, where again it increases as a step by the same amount and at every value of E_n. Density of states in the valence band behaves similarly as shown in Figure 3.35 (c). Since the electron is free in the yz plane within the well, its kinetic energy will increase parabolically and continuously with the wave vector k_{yz} in this plane as

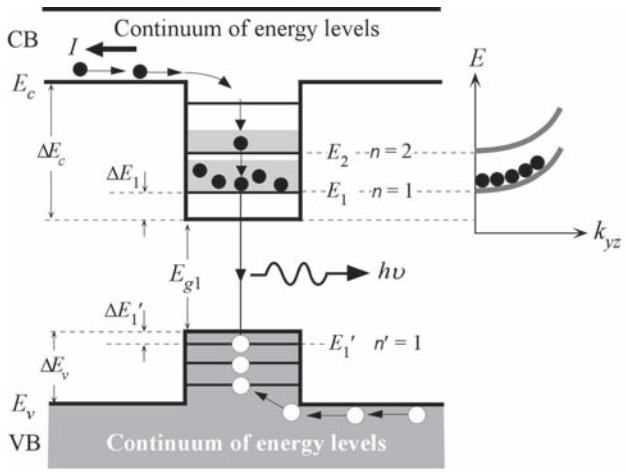


FIGURE 3.36 A QW structure that shows the energy levels in the wells and how charge carriers that are brought in by the current fall into the lowest energy level in the well and then recombine, emitting a photon. The electrons at a particular energy level also have kinetic energies in the *yz* plane, which is not quantized. The electrons are therefore spread in energy above E_n as shown. The same notion also applies to holes in the ΔE_v well.

shown in Figure 3.36. One has to account for this small kinetic energy, which is shown by shading the region above E_1, E_2, *etc.*, in Figure 3.36.

Since at E_1 there is a finite and substantial density of states, the electrons in the conduction band do not have to spread as far in energy as in the bulk to find states. In the bulk semiconductor, on the other hand, the density of states at E_c is zero and increases slowly with energy (as $E^{1/2}$), which means that the electrons are spread more deeply into the conduction band in search for states. A large concentration of electrons can easily occur near E_1 whereas this is not the case in the bulk semiconductor. Similarly, the majority of holes in the valence band will be around E_1' since there are sufficient states at this energy. Under a forward bias, as indicated in Figure 3.36, electrons are injected into the conduction band of the E_{g1}-layer, which serves as the *active layer.* The injected electrons readily populate the ample number of states at E_1 which means that the electron concentration at E_1 increases rapidly with the current. The direct recombination rate, that is, the radiative recombination rate, which determines the rate of photon emission, depends on the product *np*, the concentration of electrons *and* holes. Thus, the radiative transitions occur much more readily in the quantum well for the same current as in the bulk device. The quantum mechanical radiative recombination probability is also enhanced because the electrons and holes in the well are brought into close proximity. (However, there are still certain selection rules as there are for photon emission in the H-atom.) Furthermore, since the majority of electrons are spread around E_1 and holes around E_1', the range of emitted photon energies can be narrower than the emission spectrum in the corresponding bulk device; unless the well has been flooded with electrons at large currents as explained below. The loss of energy as the electron drops from $n = 3$ to 2 to 1, *etc.*, to lower levels inside the well occurs by the emission of phonons (lattice vibrations) and is very rapid. The transition from $n = 1$ to $n' = 1$ emits a photon of energy $h\upsilon \approx E_1 - E_1'$. (An approximate sign is used since the electron and the hole can have kinetic energies in the *yz* plane). The radiative transition in a QW must obey a **selection rule**, which requires the initial and final quantum numbers, n and n', to be the same.[22] The transition from $n = 1$ to $n' = 1$ is allowed and emits a photon, as well as that from $n = 2$ to $n' = 2$.

[22] Quantum mechanical selection rules for radiative transitions are nothing unusual; they also occur in radiative transitions in atoms. A very readable derivation can be found in M. Fox, *Optical Properties of Solids,* 2nd Edition (Oxford University Press, 2010), Ch. 6.

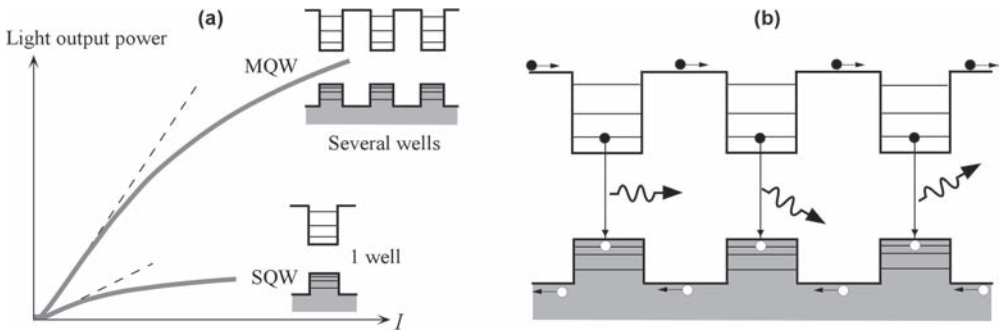


FIGURE 3.37 A schematic illustration of the comparison of light power output vs. current characteristics for an SQW and an MQW LED.

The main problem with the single quantum well (SQW) heterostructure LEDs is that, under a sufficiently large current, the well can be flooded with charge carriers and can overflow. For example, electrons can flood the QW ΔE_c and the well will overflow. The advantages of the QW action (such as confinement that increases the electron concentration n) would be lost. The light output will no longer increase proportionally to the current, and will fall behind the increase in the current as indicated in Figure 3.37 (a). This problem can be resolved by using **multiple quantum wells** (MQWs), in which electrons are shared by a number of quantum wells as in Figure 3.37 (b). Not only is the overflow problem alleviated but more photon flux is generated due to more wells as shown in Figure 3.37 (a). Modern high intensity UV, violet, and blue LEDs use MQW heterostructures. They use a thin $In_xGa_{1-x}N$ (E_{g1}) QW layer that is sandwiched between GaN (E_{g2}) layers. GaN has a large bandgap of 3.4 eV, and the composition and hence the bandgap of InGaN is chosen for the application, for example, for blue, $E_{g1} = 2.7\,\text{eV}$. The heterostructure has a number of MQWs to improve the efficiency but the number of QWs is not many, which is limited by the fabrication process.

EXAMPLE 3.12.1 Energy levels in the quantum well

Consider a GaAs QW sandwiched between two $Al_{0.40}Ga_{0.60}As$ layers. Suppose that the barrier height ΔE_c is 0.30 eV, the electron effective mass in the well is $0.067m_e$, and the width of the QW (d) is 12 nm. Calculate the energy levels E_1 and E_2 from the bottom of the well (E_c) assuming an infinite PE well as in Eq. (3.12.1). Compare these with the calculations for a *finite* PE well that give 0.022 eV, 0.088, and 0.186 for $n = 1$, 2, and 3.

Solution

We use Eq. (3.12.1) with $m_e^* = 0.067m_e$, $d = 12 \times 10^{-9}$ nm, so that for $n = 1$

$$\Delta E_n = E_n - E_c = \frac{h^2n^2}{8m_e^*d^2} = \frac{(6.624 \times 10^{-34}\,\text{J s})^2(1)^2/(1.602 \times 10^{-19}\,\text{J eV}^{-1})}{8(0.067 \times 9.1 \times 10^{-31}\,\text{kg})(12 \times 10^{-9}\,\text{m})^2} = 0.039\,\text{eV}$$

We can repeat the above calculation for $n = 2$ and 3 to find $\Delta E_2 = 0.156$ eV and $\Delta E_3 = 0.351$ eV. The third level will be above the well depth ($\Delta E_c = 0.3\,\text{eV}$). Clearly, the infinite QW predicts higher energy levels, by a factor of 1.8, and puts the third level inside the well, not outside. The finite QW calculation is not simple, and involves a numerical solution.[23]

[23] A similar example is solved for the finite well in M. Fox, *Optical Properties of Solids*, 2nd Edition (Oxford University Press, 2010), pp. 149–152. (This is a highly recommended text on optical properties.)

3.13 LED MATERIALS AND STRUCTURES

A. LED Materials

There are various direct bandgap semiconductor materials that can be readily doped to make commercial *pn* junction LEDs which emit radiation in the red and infrared range of wavelengths. The fabrication of an actual LED would require that the doped crystal layers with the required bandgap can be grown on a suitable **substrate** crystal. The substrate crystal and the LED material will have to be **lattice matched**, that is, they must have the same crystal structure and very close lattice parameters (a) to avoid creating dislocations at the interface. (Dislocations and other defects provide a nonradiative recombination path.) Lattice matching is not always perfect and often dictates what material compositions can be used.

An important class of commercial semiconductor materials that cover the visible spectrum is the **III–V ternary alloys** based on alloying GaAs and GaP, which are denoted as $GaAs_{1-y}P_y$. In this compound, As and P atoms from group V are distributed randomly at normal As sites in the GaAs crystal structure. When $y < 0.45$, the alloy $GaAs_{1-y}P_y$ is a direct bandgap semiconductor and hence the EHP recombination process is direct and efficient. The emitted wavelengths range from about 630 nm, red, for $y = 0.45$ ($GaAs_{0.55}P_{0.45}$) to 870 nm for $y = 0$, GaAs.

$GaAs_{1-y}P_y$ alloys (which includes GaP) with $y > 0.45$ are indirect bandgap semiconductors. The EHP recombination processes occur through recombination centers and involve lattice vibrations rather than photon emission. As mentioned above, however, if we add isoelectronic impurities such as nitrogen into the semiconductor crystal then the N-dopants can act as recombination centers. An electron is first captured by the N-center, the excess energy is lost to phonons, and then while at the N-center, it recombines with a hole in a radiative transition. The emitted photon energy is only slightly less than E_g. Nitrogen-doped indirect bandgap $GaAs_{1-y}P_y$ alloys are widely used in inexpensive green, yellow, and orange LEDs.

There are various commercially important direct bandgap semiconductor materials that emit in the red and infrared wavelengths which are typically **ternary** (containing three elements) and **quaternary** (four elements) alloys based on Group III and V elements, so-called **III–V alloys**. For example, GaAs with a bandgap of about 1.42 eV emits radiation at around 870 nm in the infrared. But ternary alloys based on $Al_{1-x}Ga_xAs$ where $x < 0.43$ are direct bandgap semiconductors. The composition can be varied to adjust the bandgap and hence the emitted radiation from about 640–870 nm, from deep red light to infrared.

AlGaInP is a quaternary III–V alloy (In, Ga, Al from III, and P from V) that has a direct bandgap variation with composition over the visible range. It can be lattice-matched to GaAs substrates for compositions $(Al_xGa_{1-x})_{0.5}In_{0.5}P$ where $x < 0.53$, that is, $Ga_{0.50}In_{0.50}P$ ($E_g = 1.89$ eV, red) to $Al_{0.265}Ga_{0.235}In_{0.50}P$ (2.33 eV, green). Many brands of high-intensity LEDs have been based on this material, which is likely to continue to be used in the high-intensity visible LED range, especially for the red, amber, and yellow.

The bandgap of quaternary alloys $In_{1-x}Ga_xAs_{1-y}P_y$ can be varied with composition (x and y) to span wavelengths from 870 nm (GaAs) to 3.5 μm (InAs) which includes the optical communication wavelengths of 1.3 μm and 1.55 μm. Figure 3.38 summarizes some typical wavelengths that can be emitted for a few selected semiconductor materials over the range from 0.3 μm to 1.7 μm, or from the UV to IR.

GaN is a direct bandgap semiconductor with an E_g of 3.4 eV. The blue GaN LEDs actually use the GaN alloy InGaN with a bandgap of about 2.7 eV, which corresponds to blue emission. One of the most important technological advances in the last two decades has been the development of various III-Nitride LEDs that can emit with high intensities from the UV to green. The alloys

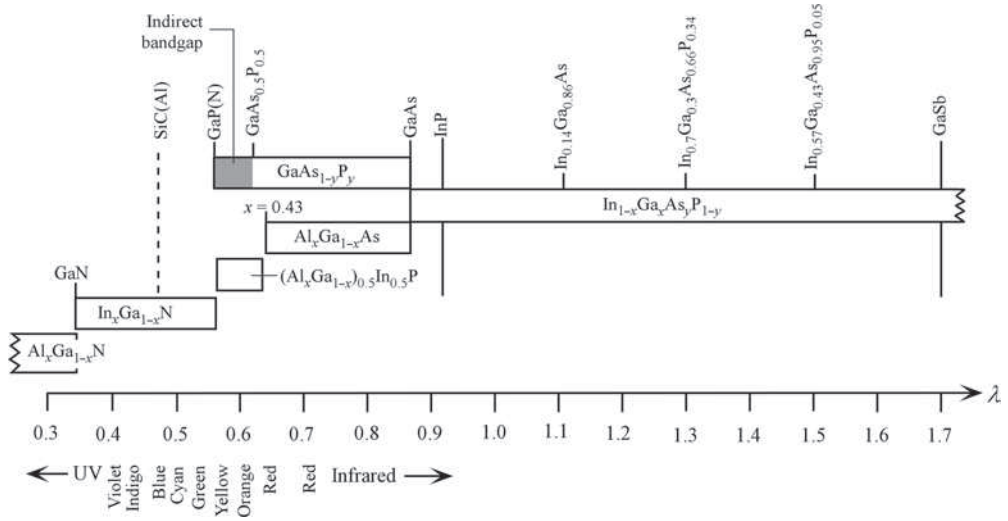


FIGURE 3.38 Free-space wavelength coverage by different LED materials from the visible spectrum to the infrared, including wavelengths used in optical communications. Grey region and dashed lines are indirect E_g materials. Only material compositions of importance have been shown.

of GaN ($E_g = 3.4\,\text{eV}$) with InN ($E_g = 0.77\,\text{eV}$), $In_xGa_{1-x}N$ spans wavelengths from the UV up to the IR, though they are currently not used beyond the green wavelength as other semiconductors such as AlGaInP provide better efficiencies. The alloys of AlN ($E_g = 6.2$ eV) and GaN ($E_g = 3.4$ eV), AlGaN, provide for wavelengths in the UV. GaN can be doped *n*-type (*e.g.*, with Si or Ge) and *p*-type (*e.g.*, with Mg), and the GaN LEDs are generally MQW heterostructures.

B. LED Structures

In its simplest technological form, LEDs are typically fabricated by *epitaxially* growing doped semiconductor layers on a suitable substrate (*e.g.*, GaAs or GaP) as illustrated for a simple homojunction *pn* junction LED in Figure 3.39 (a). This type of planar *pn* junction is formed by the epitaxial growth of first the *n*-layer and then the *p*-layer. The substrate is essentially a mechanical support for the *pn* junction device (the layers) and can be of different crystal. The *p*-side is on the surface from which light is emitted and is therefore made narrow (a few microns)

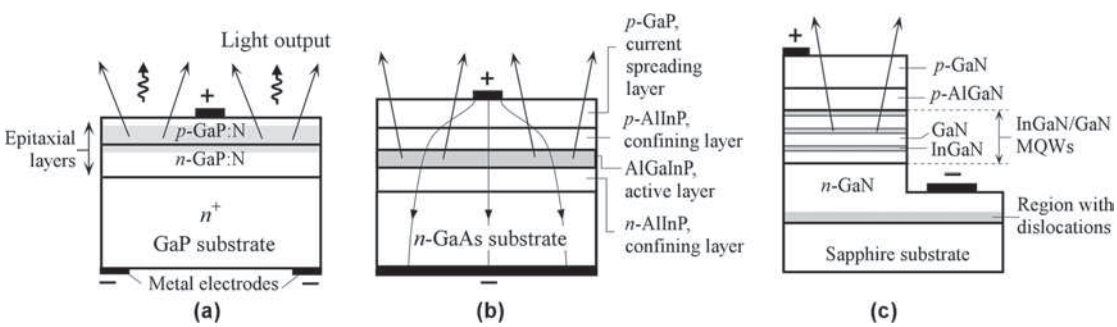


FIGURE 3.39 A schematic illustration of various typical LED structures. (a) A planar surface-emitting homojunction green GaP:N LED. (b) AlGaInP high-intensity heterostructure LED. (c) III-Nitride-based (GaN/InGaN) MQW LED for emission from the UV to green.

to allow the photons to escape without being reabsorbed. To ensure that most of the recombination takes place in the *p*-side, the *n*-side is heavily doped (n^+). Those photons that are emitted toward the *n*-side become either absorbed or reflected back at the substrate interface depending on the substrate thickness and the exact structure of the LED. The use of a segmented back electrode as in Figure 3.39 (a) will encourage reflections from the semiconductor–air interface.

If the epitaxial layer and the substrate crystals have different crystal lattice parameters, then there is a lattice mismatch between the two crystal structures. This causes lattice strain in the LED layer and hence leads to crystal defects. Such crystal defects encourage radiationless EHP recombinations. That is, a defect acts as a recombination center. Such defects are reduced by lattice matching the LED epitaxial layer to the substrate crystal. It is therefore important to lattice-match the LED layers to the substrate crystal. For example, $Al_xGa_{1-x}As$ alloys with $x < 0.43$ are direct bandgap semiconductors that have a bandgap corresponding to the IR to the red emission region (Figure 3.38). $Al_xGa_{1-x}As$ layers can be grown on GaAs substrates with excellent lattice match, which results in high-efficiency LED devices.

Figure 3.39 (b) shows the structure of a high-intensity AlGaInP heterostructure LED. The layers are grown epitaxially on an *n*-GaAs substrate. There are at least four layers. The active layer is a thin AlGaInP (*e.g.*, $Al_{0.35}Ga_{0.15}In_{0.5}P$), which is lightly doped. This layer is sandwiched by confining layers that are *p*-type and *n*-type AlInP (*e.g.*, $Al_{0.5}In_{0.5}P$) on the positive and negative terminal sides, respectively. AlInP has a wider bandgap than AlGaInP, and the band offsets confine the carriers to the active region. Remember that under forward bias the *p*-AlInP injects holes and *n*-AlInP injects electrons into the active layer. The top layer is *p*-GaP and serves to spread out the current to regions outside the top contact. Thus, radiative recombinations are avoided right under the top contact from which photons cannot be extracted. Put differently, EHP recombination is spread into regions outside the zone directly under the top contact.

Figure 3.39 (c) shows a simplified III-Nitride-based MQW LED that can be used for emission in the blue as well as green.[24] With some modification to compositions, it can also emit in the UV. The *p*-GaN (doped with Mg) is the required *p*-layer used for the injection of holes. The QWs are formed between the narrower E_g InGaN and wider E_g GaN as described above, which are undoped. There is a *p*-AlGaN layer that is called a *buffer* layer. The bandgap of AlGaN is wider than InGaAs, so it confines the injected electrons in the QW-region. The *n*-GaN layer is the electron injecting *n*-type semiconductor from which electrons are injected into the MQWs. It is difficult to find a matching crystal substrate for GaN. A sapphire crystal is the most commonly used substrate, though the mismatch is roughly 12% (significant). Special growth techniques have been developed to keep the defects (dislocations) to the initial GaN growth region near the sapphire–GaN interface, away from actual LED heterostructure. Notice that the negative terminal is on high-quality *n*-GaN, away from the defective region.

Not all light rays reaching the semiconductor–air interface, however, can escape because of total internal reflection (TIR). Those rays with angles of incidence greater than the critical angle θ_c become reflected as illustrated in Figure 3.40 (a). For the GaAs–air interface, for example, θ_c is only 17°, which means that much of the light suffers TIR. It is possible to shape the surface of the semiconductor into a dome, or hemisphere, so that light rays strike the surface at angles less than θ_c and therefore do not experience TIR. The main drawback, however, is the additional difficult

[24] III-Nitride MQW LEDs are currently the most efficient commercial LEDs in the blue, violet, and the UV. An extensive review can be found in Hadis Morkoç, *Handbook of Nitride Semiconductors and Devices*, Vol. 3 (Wiley-VCH, Weinheim, 2008), Ch. 1.

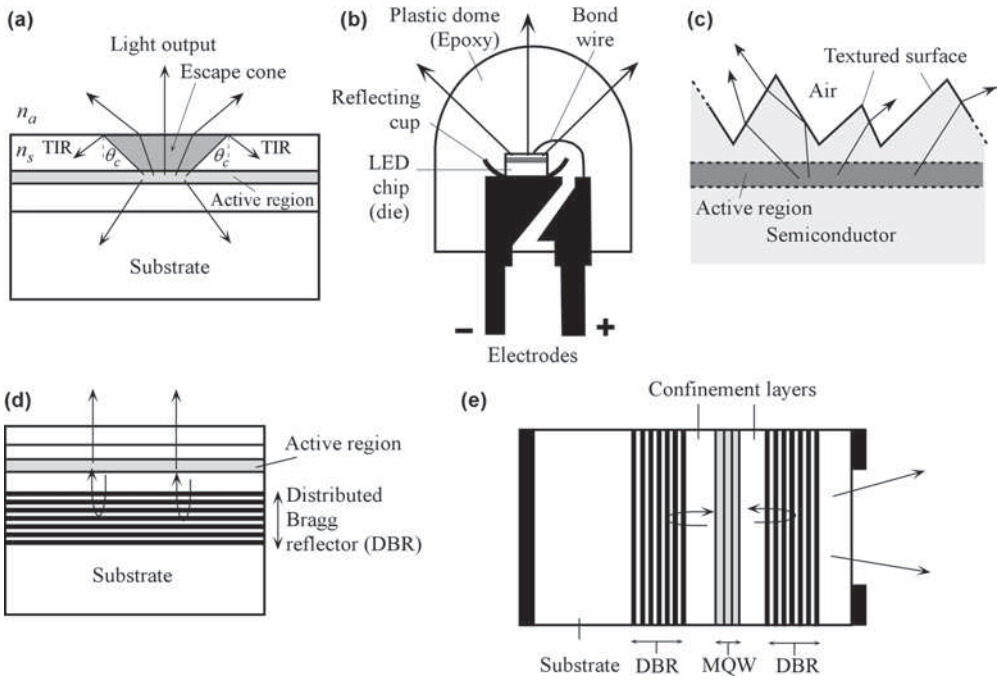


FIGURE 3.40 (a) Some of the internally generated light suffers total internal reflection (TIR) at the semiconductor–air interface and cannot be emitted into the outside. (b) A simple structure that overcomes the TIR problem by placing the LED chip at the centre of a hemispherical plastic dome. The epoxy is refractive index matched to the semiconductor and the rays reaching the dome's surface do not suffer TIR. (c) An example of a textured surface that allows light to escape after a couple of (or more) reflections (highly exaggerated sketch). (d) A distributed Bragg reflector (DBR) under the confining layer (below the active region in grey) acts as a dielectric mirror, and increases the extraction ratio. (e) An RCLED is an LED with an optical resonant cavity (RC) formed by two DBRs has a narrower emission spectrum.

process in fabricating such domed LEDs and the associated increase in expense. An inexpensive and common procedure that reduces TIR is the encapsulation of the semiconductor junction within a transparent plastic medium (an epoxy) which has a higher refractive index than air and, further, also has a domed surface on the emission side of the LED chip as shown in Figure 3.40 (b). Many individual LEDs are sold in similar types of plastic bodies.

Shuji Nakamura obtained his PhD from the University of Tokushima in Japan, and is currently a professor at the University of California at Santa Barbara and the director of Solid State Lighting and Energy Center. He has been credited with the pioneering work that has led to the development of GaN and $In_xGa_{1-x}N$-based blue and violet light–emitting diodes and laser diodes. He discovered how III-Nitrides could be doped *p*-type, which opened the way to fabricating various UV, violet, blue, and green LEDs. He is holding a blue laser diode that is turned on. *(Courtesy of Shuji Nakamura, University of California, Santa Barbara.)*

Another example of a device structure that improves the light extraction ratio is shown in Figure 3.40 (c). The surface has been textured or nanostructured. Such a textured surface allows light to escape after one or two reflections. The extraction ratio can be greatly improved if the rays traveling toward the substrate could be somehow reflected back toward the exit window. In some LEDs, a *distributed Bragg reflector* (DBR), that is a dielectric mirror, under the confining layer (below the active region) acts as a dielectric mirror at the wavelength of the LED light, and increases the extraction ratio, as shown in Figure 3.40 (d).

Another possible design is to use a *resonant cavity* for the light generated by the active region as shown in Figure 3.40 (e). An RCLED (resonant cavity LED) is an LED with an optical resonant cavity (RC) formed by two DBRs. The optical cavity is obviously wavelength-selective since only those special modes of the optical cavity that fall into the spontaneous emission spectrum can be supported or excited. The width of the emitted spectrum is significantly narrower, as much as ten times, than a conventional LED. The reason is that the light from the active region can only escape if it is a mode of the resonant cavity. The spectral width of this light is determined by the resonant cavity losses rather than the emission spectrum from the active region.

EXAMPLE 3.13.1 Light extraction from a bare LED chip

As shown in Figure 3.40 (a), due to total internal reflection (TIR) at the semiconductor–air surface, only a fraction of the emitted light can escape from the chip. The critical angle θ_c is determined by $\sin\theta_c = n_a/n_s$ where n_a and n_s are the refractive indices of the ambient (*e.g.,* for air, $n_a = 1$) and the semiconductor, respectively. The light within the **escape cone** defined by θ_c can escape into the ambient without TIR as indicated in Figure 3.40 (a). To find the fraction of light within the escape cone we need to consider solid angles, which leads to $(1/2)(1 - \cos\theta_c)$. Further, suppose that T is the average light *transmittance* of the n_s–n_a interface for those rays within the escape cone, then for a simple bare chip,[25]

$$\text{Light extraction ratio} \approx (1/2)(1 - \cos\theta_c) \times T \qquad \textbf{(3.13.1)}$$

Light extraction ratio

Estimate the extraction ratio for a GaAs chip with $n_s = 3.4$ and air as ambient ($n_a = 1$) and then with an epoxy dome with $n_a = 1.8$.

Solution

First, note that $\theta_c = \arcsin(n_a/n_s) = \arcsin(1/3.4) = 17.1°$. For T we will assume near-normal incidence (somewhat justified since the angle 17.1° is not too large) so that from Chapter 1,

$$T = 4n_s n_a/(n_s + n_a)^2 = 4(3.4)(1)/(3.4 + 1)^2 = 0.702$$

Using Eq. (3.13.1)

$$\text{Light extraction ratio} \approx (1/2)(1 - \cos\theta_c) \times T = (1/2)\left[1 - \cos(17.1°)\right] \times 0.702$$

$$\approx 0.0155 \quad \text{or} \quad 1.6\%$$

It is clear that only 1.6% of the generated light power is extracted from a bare chip, which is disappointingly small. The technological drive is therefore to improve light extraction as much as possible. If we now repeat the calculation for $n_a = 1.8$, we would find, $\theta_c = 32°$, and 6.9% light extraction.

[25] An extensive treatise on various LED structures and light extraction techniques can be found in E. F. Schubert, *Light-Emitting Diodes*, 2nd Edition (Cambridge University Press, 2006). Equation (3.13.1) is an approximation inasmuch as we should average the transmittance.

3.14 LED EFFICIENCIES AND LUMINOUS FLUX

The comparison of different LED materials and device structures requires that we define internal and external quantum efficiency. Further, we also need to meaningfully compare the output power and brightness of different light-emitting devices per unit electrical energy input.

The **internal quantum efficiency** (IQE) η_{IQE} gauges what fraction of electron hole recombinations in the forward-biased *pn* junction are radiative and therefore lead to photon emission. Nonradiative transitions are those in which an electron and a hole recombine through a recombination center such as a crystal defect or an impurity and emit phonons (lattice vibrations). Suppose that τ_r is the mean lifetime of a minority carrier before it recombines radiatively and τ_{nr} is the mean lifetime before it recombines via a recombination center (or a defect) without emitting a photon. By definition,

Internal quantum efficiency

$$\eta_{\text{IQE}} = \frac{\text{Rate of radiative recombination}}{\text{Total rate of recombination (radiative and nonradiative)}} \tag{3.14.1}$$

or

Internal quantum efficiency

$$\eta_{\text{IQE}} = \frac{\tau_r^{-1}}{\tau_r^{-1} + \tau_{nr}^{-1}} \tag{3.14.2}$$

The rate of nonradiative recombination rate $(1/\tau_{nr})$ would also include the recombination of injected carriers at the interfaces of heterostructures between different crystals: an important loss mechanism that would reduce η_{IQE}.

The total current I is determined by the total rate of recombinations whereas the number of photons emitted per second, the **photon flux** Φ_{ph}, is determined by the rate of radiative recombinations.

Internal quantum efficiency

$$\eta_{\text{IQE}} = \frac{\text{Photons emitted per second}}{\text{Total carriers lost per second}} = \frac{\Phi_{\text{ph}}}{I/e} = \frac{P_{o(\text{int})}/h\upsilon}{I/e} \tag{3.14.3}$$

where $P_{o(\text{int})}$ is the *optical* power generated internally (not yet extracted).

The **external quantum efficiency** (EQE) η_{EQE} of an LED represents the efficiency of conversion from electrical quanta, *i.e.,* electrons, that flow into the LED to optical quanta, *i.e.,* photons, that are emitted into the outside world. It incorporates the "internal" efficiency of the radiative recombination process [embedded in Eq. (3.14.1)] and the subsequent efficiency of photon extraction from the device. Suppose that the actual optical power emitted to the ambient, called the **radiant flux**, is P_o. (Φ_e is also used in the literature.) $P_o/h\upsilon$ is the number of emitted photons per second. Since the number of electrons flowing into the LED is I/e, we have

External quantum efficiency

$$\eta_{\text{EQE}} = \frac{P_o/h\upsilon}{I/e} \tag{3.14.4}$$

Thus, we should use external QE to meaningfully compare different LED efficiencies and internal QE in comparing different LED materials. For indirect bandgap semiconductors η_{EQE} are generally less than 1% whereas for direct bandgap semiconductors with the right device structure, η_{EQE} can be substantial, for example, 30–40%.

The light extraction ratio was introduced in Example 3.13.1. We can now define this more formally. The light **extraction ratio**, or the **extraction efficiency** (EE), η_{EE}, is the fraction of light that is extracted to the ambient from the internally generated light, that is,

$$\eta_{\text{EE}} = \frac{\text{Photons emitted externally from the device}}{\text{Photons generated internally by recombination}} \tag{3.14.5}$$

Extraction efficiency

Using Eqs. (3.14.1), (3.14.3) and (3.14.5), the emitted optical output power, the radiant flux, is

$$P_o = \eta_{\text{EE}} P_{o(\text{int})} = h\upsilon \eta_{\text{EE}} \eta_{\text{IQE}} (I/e) \tag{3.14.6}$$

Emitted optical power

The **power conversion efficiency** (PCE), η_{PCE}, or simply the **power efficiency**, gauges the overall efficiency of conversion from the input of electrical power to the output of optical power, *i.e.,*

$$\eta_{\text{PCE}} = \frac{\text{Optical output power}}{\text{Electrical input power}} = \frac{P_o}{IV} \approx \eta_{\text{EQE}} \left(\frac{E_g}{eV} \right) \tag{3.14.7}$$

Power efficiency

If we wish to compare the brightness of light-emitting devices per unit electrical power input then we need to examine the luminous flux emitted. The visual "brightness" of a source as observed by an average daylight-adapted eye is proportional to the radiation (optical) power emitted, that is, the radiant flux, and the efficiency of the eye to detect the spectrum of the emitted radiation. While the eye can see a red color source, it cannot see an infrared source and the brightness of the infrared source would be zero. The **luminous flux** Φ_v is a measure of *visual* **brightness**, in lumens (lm), and is defined by

$$\Phi_v = P_o \times (633 \text{ lm W}^{-1}) \times V(\lambda) \tag{3.14.8}$$

Luminous flux

where P_o is the radiant flux or the radiation power emitted (in watts) and $V(\lambda)$ is the **relative luminous efficiency** (or the relative sensitivity) of an average light-adapted (photopic) eye, which depends on the wavelength and hence λ in parenthesis. The function $V(\lambda)$ is also called the **luminosity function** and the **visibility function**. $V(\lambda)$ is a Gaussian-like function with a peak of unity at 555 nm as shown in Figure 3.41. One lumen of luminous flux, or brightness, is obtained from a 1.58 mW light source emitting at a single wavelength of 555 nm (green). A typical 60 W incandescent lamp provides roughly 900 lm (or 15 lm W^{-1}). When we buy a light bulb, we are buying lumens.

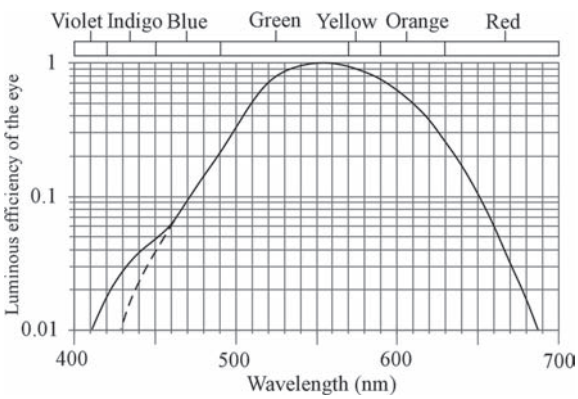


FIGURE 3.41 The luminous efficiency $V(\lambda)$ of the light-adapted (photopic) eye as a function of wavelength. The solid curve is the Judd-Vos modification of the CIE 1924 photopic photosensitivity curve of the eye. The dashed line shows the modified region of the original CIE 1924 curve to account for its deficiency in the blue-violet region. (The vertical axis is logarithmic.)

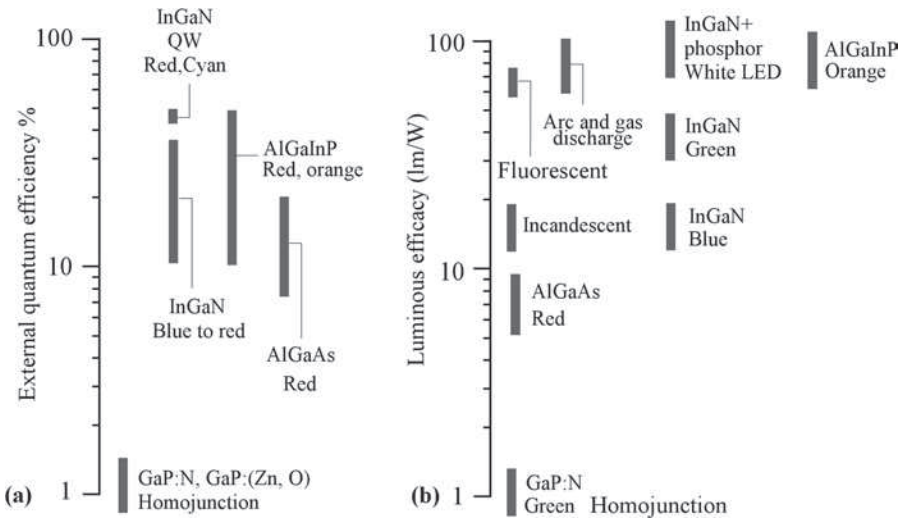


FIGURE 3.42 Typical (a) external quantum efficiency and (b) luminous efficacy of various selected LEDs, and how they stand against other light sources such as the fluorescent tube, arc, and gas discharge lamps and the incandescent lamp.

The **luminous efficacy**[26] **of a light source** (such as a lamp) as commonly used in *lighting applications* is the efficiency with which an electrical light source converts the input electric power (watts) into an emitted luminous flux (lumens).

Luminous efficacy (efficiency)

$$\eta_{\text{LE}} = \frac{\Phi_v}{IV} \tag{3.14.9}$$

A 100 W light bulb producing 1700 lumens has an efficacy of 17 lumens watt^{-1} (lm W^{-1}). Recent technological advances have led to LEDs with efficacies that are comparabale to standard fluorescent tubes—efficacies around 100 lm W^{-1}. LEDs as solid state lamps have much longer lifetimes and much higher reliability, and hence are expected to be more economical than incandescent and fluorescent lamps. Figure 3.42 (a) and (b) show some typical external efficiency and luminous efficacy values of various selected LEDs. The advantages of III-Nitrides (*e.g.*, InGaN) are clearly obvious. It is also important to note that in some cases, as in InGaN based LEDs, the high EQE and hence luminous efficacy cannot be maintained as the current is increased to obtain higher luminous flux (more lumens) because the output flux is not directly proportional to the input current, especially at high currents, as apparent in Figure 3.37 (a).

EXAMPLE 3.14.1 LED efficiencies

A particular 870 nm IR LED for use in optical links and instrumentation has a GaAs chip. Active layer that has been doped p-type with $2 \times 10^{17}\,\text{cm}^{-3}$ of acceptors and the nonradiative lifetime is about 100 ns. At a forward current of 30 mA, the voltage across it is 1.35 V, and the emitted optical power is 6.5 mW. Calculate the IQE, EQE, and PCE, and estimate the light extraction ratio. For GaAs, $B \approx 2 \times 10^{-16}\,\text{m}^2\,\text{s}^{-1}$.

[26] Some authors use the term *luminous efficiency* but the latter, strictly, needs the output and input quantities to have the same units so that the efficiency can be expressed as a percentage, which is not the case here. *Efficacy* would be a better term.

Solution

The radiative lifetime $\tau_r = 1/BN_a = 1/\left[(2 \times 10^{-16}\,\mathrm{m^3\,s^{-1}})(2 \times 10^{23}\,\mathrm{m^{-23}})\right] = 2.5 \times 10^{-8}\,\mathrm{s}$ or 25 ns. IQE is,

$$\eta_{\mathrm{IQE}} = \frac{\tau_r^{-1}}{\tau_r^{-1} + \tau_{nr}^{-1}} = \frac{(25\,\mathrm{ns})^{-1}}{(25\,\mathrm{ns})^{-1} + (100\,\mathrm{ns})^{-1}} = 0.80 \quad \text{or} \quad 80\%$$

The emitted photon energy $h\upsilon = hc/\lambda \approx 1.43$ eV. Thus the EQE is

$$\begin{aligned}\eta_{\mathrm{EQE}} &= (P_o/h\upsilon)/(I/e)\\ &= \left[(6.5 \times 10^{-3}\,\mathrm{W})/(1.43\,\mathrm{eV} \times 1.6 \times 10^{-19}\,\mathrm{J\,eV^{-1}})\right]/\left[(30 \times 10^{-3}\,\mathrm{A})/(1.6 \times 10^{-19}\,\mathrm{C})\right]\\ &= 0.15 \quad \text{or} \quad 15\%\end{aligned}$$

The PCE is simply P_o/IV or $(6.5\,\mathrm{mW})/\left[(30\,\mathrm{mA})(1.35\,\mathrm{V})\right]$, that is, 0.16, *i.e.*, 16%.
From Eq. (3.14.6), using $P_o = h\upsilon\eta_{\mathrm{EE}}\eta_{\mathrm{IQE}}(I/e)$,

$$6.5 \times 10^{-3}\,\mathrm{W} = (1.43\,\mathrm{eV} \times 1.6 \times 10^{-19}\,\mathrm{J\,eV^{-1}})\eta_{\mathrm{EE}}(0.80)(30 \times 10^{-3}\,\mathrm{A}/1.6 \times 10^{-19}\,\mathrm{C})$$

solving the above gives $\eta_{\mathrm{EE}} = 0.19$ or 19%.

EXAMPLE 3.14.2 LED brightness

Consider two LEDs, one red, with an optical output power (radiant flux) of 10 mW, emitting at 650 nm, and the other, a weaker 5 mW green LED, emitting at 532 nm. Find the luminous flux emitted by each LED.

Solution

For the red LED, at $\lambda = 650\,\mathrm{nm}$, Figure 3.41 gives $V \approx 0.10$ so that from Eq. (3.14.8)

$$\Phi_v = P_o \times (633\,\mathrm{lm\,W^{-1}}) \times V = (10 \times 10^{-3}\,\mathrm{W})(633\,\mathrm{lm\,W^{-1}})(0.10) = 0.63\,\mathrm{lm}$$

For the green LED, $\lambda = 532$ nm, Figure 3.41 gives $V \approx 0.87$ so that from Eq. (3.14.8)

$$\Phi_v = P_o \times (633\,\mathrm{lm\,W^{-1}}) \times V = (5 \times 10^{-3}\,\mathrm{W})(633\,\mathrm{lm\,W^{-1}})(0.87) = 2.8\,\mathrm{lm}$$

Clearly the green LED at half the optical power is 4 times brighter than the red LED.

3.15 BASIC LED CHARACTERISTICS

The *I–V* characteristics of LEDs depend on the device structure, including the material properties of the semiconductors used in the LED. The DC current generally increases steeply with the voltage, as shown in Figure 3.43 (a). The *I–V* characteristics do not always follow an exact exponential behavior that is often seen for a simple forward-biased *pn* junction. Most modern LEDs are heterostructure devices, with several layers of differently doped semiconductors. The current increases sharply over a narrow voltage range, as in Figure 3.43 (a), where it can be seen that there is an apparent **turn-on** or a **cut-in voltage** beyond which the current increases very sharply with voltage. Most LED manufacturers, however, quote the **forward voltage** V_F when the LED is operating fully; for example, the manufacturer of one type of red AlGaInP-based LED quotes $V_F = 2.0\,\mathrm{V}$ at $I = 20$ mA, at 40% rated maximum current. V_F depends on both the semiconductor material and the device structure. There is an overall trend in which V_F tends to increase with decreasing wavelength, that is, with increasing photon energy or roughly the bandgap E_g of the active region. However, the device structure is also very important, and can skew this

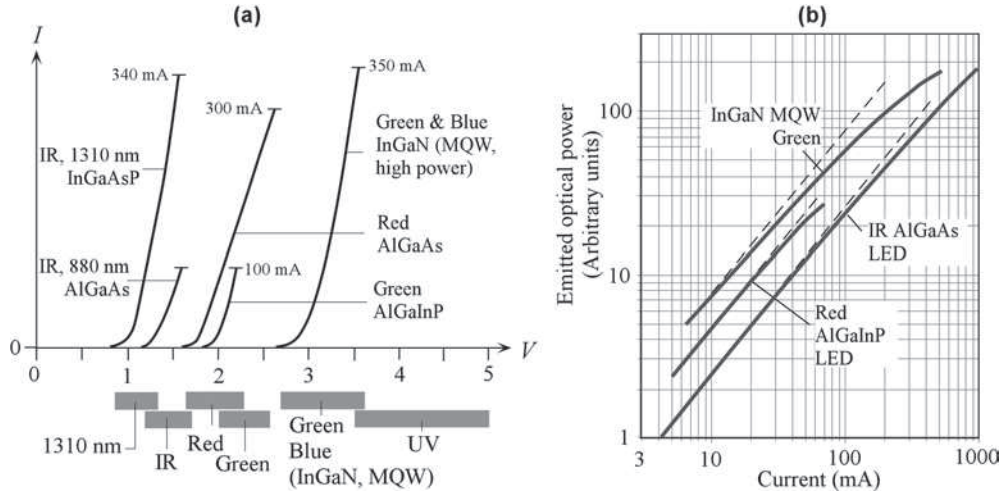


FIGURE 3.43 (a) Current-voltage characteristics of a few LEDs emitting at different wavelengths from the IR to blue. (b) Log-log plot of the emitted optical output power vs. DC current for three commercial devices emitting at IR (890 nm), red and green. The vertical scale is in arbitrary units and the curves have been shifted to show the dependence of P_o on I. The ideal linear behavior $P_o \propto I$ is also shown.

observation. MQW InGaN LEDs have a higher V_F than corresponding AlGaInP heterojunction LEDs, that is, they operate at a higher voltage across the device as apparent for the two green LEDs in Figure 3.43 (a). Indeed, InGaN green to violet LEDs all have very similar V_F.

Typical optical output power (P_o) vs. current (I) characteristics of LEDs are shown in Figure 3.43 (b) on a log-log plot for three cases. For comparison, the expected linear relationship, $P_o \propto I$, has been also shown for each device. In general, at high currents, P_o vs. I relationship curves down from the expected linear, $P_o \propto I$, behavior. The worst case is for InGaN MQW LEDs in which there is significant deviation from the expected linear relationship almost from the start, that is, P_o cannot keep up linearly with the current and droops as the current increases. This droop represents a drop in the quantum efficiency of the MQW InGaN device, and is currently a topical area of research. The P_o-I characteristics for standard AlGaInP and AlGaAs heterojunction LEDs deviate from linearity mainly at high currents, exhibiting an extensive range of reasonable linearly. While a non-linear P_o-I behavior is not a serious problem in digital communications, it can create distortion in analog modulation, especially under large signals.

3.16 LEDS FOR OPTICAL FIBER COMMUNICATIONS

The type of light source suitable for optical communications depends not only on the communication distance but also on the bandwidth requirement. For short-haul applications, for example, local networks, LEDs are preferred as they are simpler to drive, more economic, have a longer lifetime, and provide the necessary output power even though their output spectrum is much wider than that of a laser diode. LEDs are frequently used with graded index fibers inasmuch as typically dispersion in a graded index fiber is primarily due to intermodal dispersion rather than intramodal dispersion. (For example, at 1310 nm communications, the material and hence chromatic dispersion in a graded index fiber is usually much smaller than intermodal dispersion.) For long-haul and wide-bandwidth communications, invariably laser diodes are used because of their narrow linewidth, high output power, and higher signal bandwidth capability.

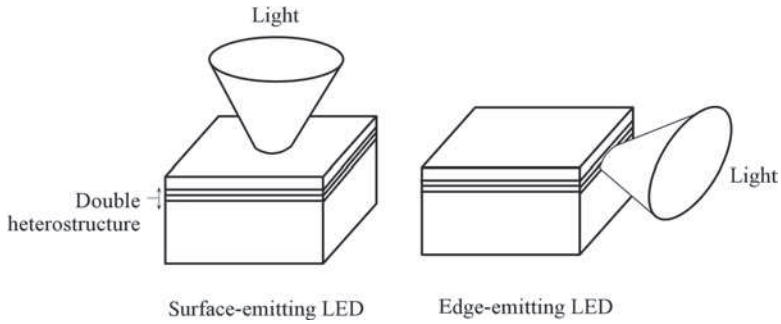


FIGURE 3.44
A surface-emitting and an edge-emitting LED.

There are essentially two types of LED devices which are illustrated in Figure 3.44. If the emitted radiation emerges from an area in the plane of the recombination layer as in (a) then the device is a **surface-emitting LED** (SLED). If the emitted radiation emerges from an area on an edge of the crystal as in (b), that is, from an area on a crystal face perpendicular to the active layer, then the LED is an **edge-emitting LED** (ELED).

The simplest method of coupling the radiation from a surface-emitting LED into an optical fiber is to etch a well in the planar LED structure and lower the fiber into the well as close as possible to the active region where emission occurs. This type of structure, as shown in Figure 3.45 (a), is called a **Burrus-**type device (after its originator). An epoxy resin is used to bond the fiber and provide refractive index matching between the glass fiber and the LED material to capture as much of the light rays as possible. Note that in the double heterostructure LED used in this way, the photons emitted from the active region, with a smaller bandgap, do not get absorbed by the neighboring layer, which has a wider bandgap. Another method is to use a truncated **spherical lens** (a microlens) with a high refractive index ($n = 1.9 - 2$) to focus the light into the fiber as shown in Figure 3.45 (b). The lens is bonded to the LED with a refractive index–matching cement and, in addition, the fiber can be bonded to the lens with a similar cement.

Edge-emitting LEDs provide a greater intensity light and also a beam that is more collimated than the surface-emitting LEDs. Figure 3.46 shows the structure of a typical edge-emitting LED for operation at ~1.5 μm. The light is guided to the edge of the crystal by a **dielectric waveguide** formed by wider bandgap semiconductors surrounding a double heterostructure. The recombination of injected carriers occurs in the InGaAs active region which has a bandgap $E_g \approx 0.83\,\text{eV}$. Recombination is confined to this layer because the surrounding InGaAsP layers, **confining layers**, have a wider bandgap ($E_g \approx 1\,\text{eV}$) and the InGaAsP/InGaAs/InGaAsP layers form a

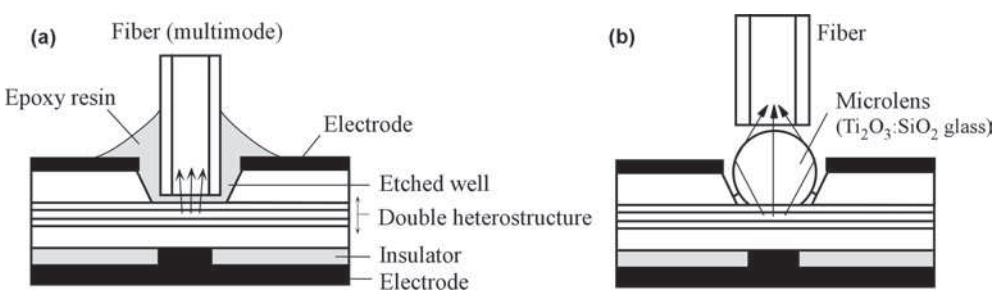


FIGURE 3.45 Coupling of light from surface-emitting LEDs into optical fibers. (a) Light is coupled from a surface-emitting LED into a multimode fiber using an index matching epoxy. The fiber is bonded to the LED structure. (b) A microlens focuses diverging light from a surface-emitting LED into a multimode optical fiber.

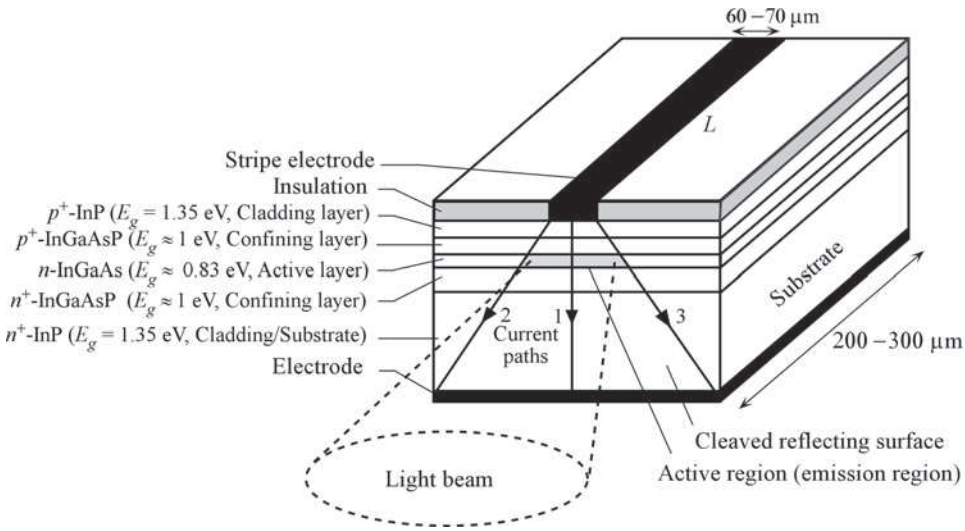


FIGURE 3.46 Schematic illustration of the structure of a double heterojunction stripe contact edge-emitting LED. (Upper case notation for a wider bandgap semiconductor is not used as there are several layers with different bandgaps.)

double heterostructure. The light emitted in the active region (InGaAs) spreads into the neighboring layers (InGaAsP) which guide it along the crystal to the edge. InP has a wider bandgap ($E_g \approx 1.35\,\text{eV}$) and thus a lower refractive index than InGaAsP. The two InP layers adjoining the InGaAsP layers therefore act as **cladding layers** and thereby confine the light to the DH structure.

Generally some kind of lens system is used to conveniently couple the emitted radiation from an ELED into a fiber. For example, in Figure 3.47 (a), a hemispherical lens attached to the fiber end is used for collimating the beam into the fiber. **A graded index (GRIN) rod lens** is a glass rod that has a parabolic refractive index profile across its cross-section with the maximum index on the rod axis. It is like a large diameter short length graded index "fiber" (typical diameters are 0.5–2 mm). A GRIN rod lens can be used to focus the light from an ELED into a fiber as illustrated in Figure 3.47 (b). This coupling is particularly useful for single mode fibers inasmuch as their core diameters are typically ~10 μm.

The output spectra from surface and edge-emitting LEDs using the same semiconductor material are not necessarily the same. The first reason is that the active layers have different doping levels. The second is the self-absorption of some of the photons guided along the active layer as in the ELED. Typically the linewidth of the output spectrum from an ELED is *less* than that from a SLED. For example, in a particular set of experiments on an InGaAsP ELED operating near 1300 nm, the emission linewidth was reported as 75 nm whereas the corresponding SLED at the same wavelength had a linewidth of 125 nm, which is significantly wider.

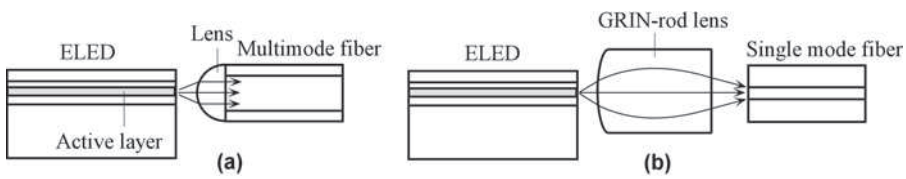


FIGURE 3.47 Light from an edge-emitting LED is coupled into a fiber typically by using a lens or a GRIN rod lens.

3.17 PHOSPHORS AND WHITE LEDS

Photoluminescence is the emission of light by a material, called a **phosphor**, that has been first excited by light of higher frequency; higher energy photons are first absorbed, and then lower energy photons are emitted. Typically the emission of light occurs from certain dopants, impurities, or even defects, called luminescent or **luminescence centers**, purposefully introduced into a **host matrix**, which may be a crystal or glass. The luminescent center is also called an **activator**. Many phosphors are based on activators doped into a host matrix; for example, Eu^{3+} (europium ion) in a Y_2O_3 (yttrium oxide) matrix is a widely used modern phosphor. When excited by UV radiation, it provides an efficient luminescence emission in the red (around 613 nm). It is used as the red-emitting phosphor in color TV tubes and in modern tricolor fluorescent lamps. Another important phosphor is Ce^{3+} in $Y_3Al_5O_{12}$ (YAG), written as $Y_3Al_5O_{12}$:Ce^{3+}, which is used in white LEDs. YAG:Ce^{3+} can absorb blue radiation and emit yellow light.

In very general terms, we can represent the energy of an activator in a host matrix by the highly simplified energy diagram in Figure 3.48 (a). We can take this figure to very roughly represent the energy of Ce^{3+} in YAG. Although it looks complicated, there are actually two manifolds of energies involved in blue absorption and yellow emission. There is a manifold of energies, labeled E_1, starting at E_1', which represents the ground energy of the activator. There is a higher manifold of energies, labeled E_2, starting at E_2'. A manifold such as E_1 has a number of energy levels that are closely spaced. The energy levels within the E_1- and E_2-manifolds represent the vibrational energies of the Ce^{3+}-ion in the host (YAG).[27] The slight horizontal shifts in the energy levels in a manifold in Figure 3.48 (a) indicate the relative position (or the tiny little displacement) of the Ce^{3+} ion in the host. Radiative transitions (absorption or emission of a photon) occur quickly compared with the time scale of vibrations of the Ce^{3+}-ion so that, during the radiative transition, the Ce^{3+} ion is essentially stationary. Thus, these radiative transitions are vertical lines in the activator energy diagram in Figure 3.48 (a); this rule is called the **Franck–Condon principle**.

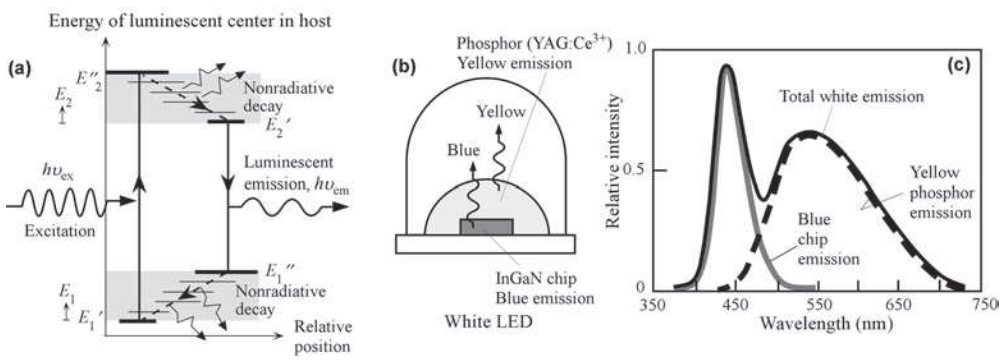


FIGURE 3.48 (a) A simplified energy diagram to explain the principle of photoluminescence. The activator is pumped from E_1' to E_2''. It decays nonradiately down to E_2'. The transition from E_2' down to E_1'' emits a photon with a lower energy than the excitation photon. (b) Schematic structure of a blue chip yellow phosphor white LED (c) The spectral distribution of light emitted by a white LED. Blue luminescence is emitted by GaInN chip and "yellow" phosphorescence is produced by phosphor. The combined spectrum looks "white."

[27] The energy diagram in Figure 3.48 (a) is still oversimplified. The vibrations of the Ce^{3+} ion is quantized and these energy levels are added to the usual electronic energy levels. However, there is also a dependence on the position of the Ce^{3+} ion, which has been simplified and represented as slightly shifted energy levels along the horizontal direction.

Upon excitation by an incident radiation of suitable energy $h\nu_{ex}$, the activator becomes excited to E_2'' in the E_2-manifold; a vertical transition. From this energy level, it decays, or *relaxes*, down relatively quickly (on a time scale of the order of picoseconds) to an energy level E_2' by emitting phonons or lattice vibrations. This type of decay is called **radiationless** or **non-radiative decay**. The Ce^{3+} ion becomes slightly displaced as shown in Figure 3.48 (a). From E_2', the activator decays down to E_1' in the E_1-manifold by emitting a photon by spontaneous emission, which is the emitted luminescent radiation. This is a vertical transition (Franck–Condon principle). The emitted photon energy is $h\nu_{em}$, which is less than the excitation photon energy $h\nu_{ex}$. The return from E_1' to the ground E_1 state involves phonon emissions. In some activators, the higher levels may form multilevel narrow energy "bands." In this example, the activator absorbed the incident radiation, and was directly excited, which is known as **activator excitation**. The Ce^{3+} ions in $Y_2Al_5O_{12}:Ce^{3+}$ can be excited directly by blue light, and would then emit in the yellow region as discussed below. It is apparent from Figure 3.48 (a) that the emitted radiation ($h\nu_{em}$) has a *longer* wavelength than the exciting radiation ($h\nu_{ex}$), that is, $h\nu_{em} < h\nu_{ex}$. The downshift in the light frequency from absorbed to emitted radiation is called the **Stokes shift**. It should be emphasized that the energy levels of the activator also depend on the host, because the internal electric fields within the host crystal act on the activator, and shift these levels up and down. The emission characteristics depend firstly on the activator, and secondly on the host.

Commercially available and popular white LEDs can provide bright light at a fraction of input electric power compared with conventional tungsten-based incandescent lights. In addition, the white LED has an extremely long lifetime. Most common white LEDs use a chip emitting at a short wavelength (blue or violet) and a *phosphor*, which absorbs some of the light from the blue diode and undergoes secondary luminescence emissions at a longer wavelength (yellow) as illustrated in Figure 3.48 (b). The mixture of blue light from the chip and yellow light from the phosphor results in an overall spectrum, shown in Figure 3.48 (c), that appears white to the eye. The quality and spectral characteristics of the combined emission in Figure 3.48 (c) vary with different white LED designs, for example, chip and phosphor combination.

Typically, a phosphor is composed of an inorganic host substance that contains optically active dopants; the dopants are the atoms that emit the lower energy photons, that is, the activators. Yttrium aluminum garnet (YAG) is a common host material. For white LED applications, it is usually doped with one of the rare-earth elements or a rare-earth compound. Cerium (Ce) is a common dopant element in YAG phosphors designed for white light–emitting diodes in

LUXEON Rebel ES white-emitting LED. *(Courtesy of Philips Lumileds.)*

Snap LED emitting in the amber for automotive signaling applications. The lamp includes the driver under the LED. *(Courtesy of Philips Lumileds.)*

The Audi A8 uses LEDs for nearly all its lighting, including headlights. *(Used with permission of Audi of America, Inc.)*

so-called **blue chip and yellow phosphor white LEDs**. White LEDs are now challenging the existing incandescent sources for general lighting.

In applications requiring a full spectrum of colors from a single point source, red, green, and blue (RGB)–emitting chips in a single package are used, which allows the generation of white light or any of 256 colors by utilizing circuits that drive the three diodes independently.

Additional Topics

3.18 LED ELECTRONICS

The electronics involved in driving LEDs depends on the application, which generally falls into three categories: (a) DC operation, (b) modulation of the LED output light intensity about some DC value in analog applications, and (c) pulsed operation in digital applications, including optical communications. In DC operation, the LED is driven by providing the necessary current for the required optical output. The *I–V* characteristics is such that large changes in the current correspond to small voltage changes across the LED as in Figure 3.43 (a), and for a given LED, V_F does not vary very much. The simplest circuit is shown in Figure 3.49 (a), where a DC voltage $V(>V_F)$ provides the current, which is determined by $I_F = (V - V_F)/R$. The problem is that keeping I_F constant depends on keeping V constant. Most LEDs have low reverse breakdown voltages and usually another diode, such as D, is used in parallel to shunt the LED when a reverse voltage is accidently applied to the LED.

A better circuit than that in Figure 3.49 (a) should provide the necessary drive current I_F to the LED, with some convenient way of controlling I_F. Bipolar junction transistors (BJTs) are commonly used in many applications as convenient current sources. Figure 3.49 (b) is shown as an example in which the LED is in the collector circuit of the BJT, which is driven by the output of an op amp A. The resistor R_F in the emitter of the BJT provides feedback, $I_F R_F$, to the negative input of A, which stabilizes the circuit. It is straightforward to show that $I_F \approx V/R_F$, that is, the circuit provides a linear control of I_F by the input voltage V. There are many ICs available that will perform the function of providing the necessary constant current under a wide range of input

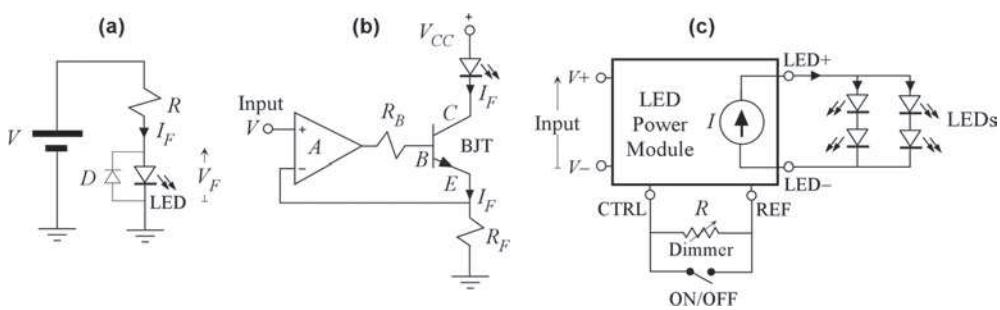


FIGURE 3.49 (a) The simplest circuit to drive an LED involves connecting it to a voltage supply (V) through a resistor R. (b). Bipolar junction transistors are well suited for supplying a constant current. Using an IC and negative feedback, the current is linearly controlled by V. (c) There are various commercial LED driver modules that can be easily configured to drive a number of LEDs in parallel and/or series. The example has a module driving four LEDs, a dimmer (R), and an on/off switch.

voltages, and temperature variations. Without a proper control of the current, the temperature variations can cause substantial changes in the current and hence the light output. Figure 3.49 (c) uses one of several LED driver modules available in the market, which can drive more than one LED in various parallel and series combinations, provided the maximum current rating is satisfied. The current can be externally and easily controlled, for example, by an external variable resistor, or a microprocessor.

The LED drive circuit shown in Figure 3.49 (b) can also be used in modulating the output of the LED as well since the current is proportional to the input voltage as in Figure 3.50 (a). One can, of course, design faster drive circuits, but the bottleneck will eventually be in the speed of response of the LED. An AC signal of frequency f on top of a DC voltage is used to drive the LED in Figure 3.50 (a). The frequency dependence of the output light power $P_o(f)$ with respect to its DC value $P_o(0)$ is shown in Figure 3.50 (b). There is a clear roll-off at high frequencies. The **cutoff frequency** f_c is defined as the frequency at which the normalized power drops by a factor of $2^{1/2}$, that is, at $f = f_c$, $P_o(f_c) = 0.707P_o(0)$. Any modulation of the LED current would result in modulating the minority carrier injection profile as shown in Figure 3.24 (a). The small signal equivalent circuit shown in Figure 3.24 (c) is usually sufficient to understand the frequency response of the LED but we have to be careful inasmuch as r_d and C_{diff} also depend on the frequency at high frequencies. First, consider very low frequencies, that is, almost DC. The capacitances are negligible. The diode action embedded in $I = I_o \exp(eV/k_BT)$ and hence the optical power output P_o, depend on the current in r_d. At high frequencies, the capacitances shunt away this current from r_d and decrease the optical output power. The characteristic time constant τ should be r_dC_{diff} (C_{diff} is greater than the depletion layer capacitance C_{dep}), that is, the net minority carrier recombination time. A more rigorous treatise arrives at the same conclusion and gives the output optical power $P_o(\omega)$ at frequency ω as

LED optical power output

$$P_o(\omega)/P_o(0) = 1/\left[1 + (\omega\tau)^2\right]^{1/2} \qquad \textbf{(3.18.1)}$$

At the critical frequency $f_c = 1/(2\pi\tau)$ in Eq. (3.18.1), the relative output power is 0.707 as shown in Figure 3.50 (b). The **optical bandwidth** f_{op} is defined as the frequency at which $P_o(f_{\text{op}})/P_o(0) = 1/2$, and occurs when $f_{\text{op}} = \sqrt{3}f_c$, that is,

Optical bandwidth of an LED

$$f_{\text{op}} = \sqrt{3}/(2\pi\tau) \qquad \textbf{(3.18.2)}$$

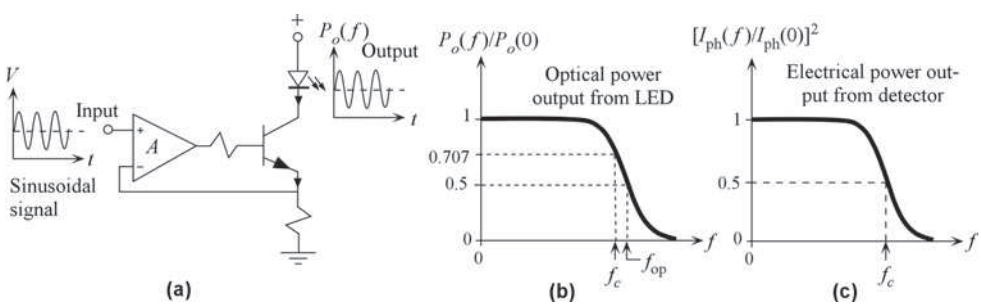


FIGURE 3.50 (a) Sinusoidal modulation of an LED. (b) The frequency response where f_c is the cutoff frequency at which $P_o(f)/P_o(0)$ is 0.707. (c) The electrical power output from the detector as a function of frequency. At f_c, $\left[I_{\text{ph}}(f)/I_{\text{ph}}(0)\right]^2$ is 0.5. However, it is 0.707 at a lower frequency than f_c.

where τ is the effective or net recombination time that is, $1/\tau = 1/\tau_r + \tau_{nr}$, where τ_r and τ_{nr} are the radiative and nonradiative recombination lifetimes. Recombination centers in the semiconductor reduce τ_{nr} and hence τ, which would widen the bandwidth. However, the IQE is also reduced, so that the optical output power is diminished as well, which is normally undesirable. If we feed the output from the LED to a photodetector, then the detector photocurrent I_{ph} will be proportional to P_o and the electrical power delivered by the detector will be proportional to I_{ph}^2 or P_o^2. At the frequency f_c, the detector's electrical output power will be proportional to $[0.707P_o(0)]^2$ or $0.5P_o(0)^2$. Thus, f_c represents a cutoff frequency (a bandwidth) at which the electrical power available from the detector has fallen to half its unmodulated value ($f = 0$ value), as illustrated in Figure 3.50 (c); f_c is called the **electrical bandwidth**.

The turning on and off of an LED in digital applications involves pumping the minority carriers into the active region and then removing them when the LED is turned off. In principle, the minority carrier lifetime τ should also play a key role in this pulsed operation. Figure 3.51 shows a simplified diagram for driving the LED digitally, for example, from a logic gate. The light output pulse should have an exponential rise and fall portions that delay the full turn-on and turn-off of the LED. The characteristic time constant for the rise and fall portions of the pulse should be the net recombination time τ. In practice, exponential rise and fall only serve as approximations to the switching behavior of the output light pulse. The net recombination time τ itself can depend on the current, and hence we should not expect a perfect exponential rise and fall.

Most LED data sheets quote the rise and fall times τ_R and τ_F for LEDs that have been pulsed. τ_R is the rise time from 10% to 90% of the final pulse height as shown in Figure 3.51. Similarly the fall time τ_F is the time it takes for the pulse to drop from its 90% to 10% value before the turn-off was triggered. If we approximate the rise and fall portions as exponentials as mentioned above, then, $\tau_R = \tau_F = 2.2\tau$. It is possible to shape the output pulse by intruding a pulse shaping capacitor C across the resistor R as shown in Figure 3.51. Upon triggering, the capacitor can provide the initial large current to pump (inject) the minority carriers in and hence turn on the device more quickly. Similarly, the capacitor would allow a large turn-off current to take out the minority carriers. Thus τ_R and τ_F can be shortened. The RC time constant must match the net recombination time in the LED for best reduction in the rise and fall times.

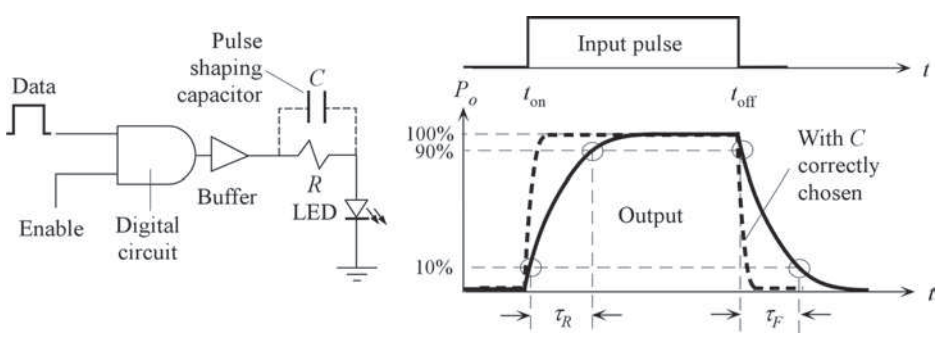


FIGURE 3.51 An LED in a digital circuit is turned on and off by a logic gate, assumed to have a buffered output as shown, to avoid being loaded. A BJT can be used after the logic gate to drive the LED as well (not shown). Definitions of rise and fall times are shown in the light output pulse.

Questions and Problems

3.1 Metals and work function The metal sodium (Na) has an atomic concentration of 2.54×10^{22} cm^{-3}. Each Na atom in isolation has one outer valence electron in an unfilled 3*s*-subshell. Once the solid is formed, the outer valence electrons from all Na atoms are shared in the crystal by all Na^+ ions, that is, these valence electrons find themselves in an unfilled energy band as in Figure 3.1 (c). Since each Na atom donates one electron to the energy band, the electron concentration is the same as the atomic concentration. Calculate the Fermi energy at 0 K. What is the speed of the electrons at E_F? What should be the speed at 300 K if the electrons could be treated classically in terms of the kinetic molecular theory as if they were free, similar to the atoms in a gas?

3.2 Photocathode and work function The photocathode of a photomultiplier tube has a multi-alkaline (Sb-Na-K-Cs) metal with a work function (Φ) of 1.55 eV. What is the longest wavelength that will cause photoemission? What is the kinetic energy of a photoemitted electron if the incident light wavelength is 450 nm (blue)? The quantum efficiency (QE) of a photocathode is defined by

$$\text{Quantum efficiency} = \frac{\text{Number of photoemitted electrons}}{\text{Number of incident photons}}$$

The QE is 100% if each incident photon ejects one electron. Suppose that blue light of wavelength 450 nm with an intensity of 1 μW cm^{-2} is incident on this photocathode with an area of 50 mm^2. If the emitted electrons are collected by applying a positive bias voltage to an anode, and the photocathode has a QE of 25%, what will be the photocurrent? (Normally the photoemitted electron is accelerated by a suitable applied field and impacts another electrode, a dynode, where it causes secondary electron emission, and so on, until the current is multiplied by orders of magnitude.)

3.3 Refractive index and bandgap Diamond, silicon, and germanium all have the same diamond unit cell. All three are covalently bonded solids. Their refractive indices (n) and energy bandgaps (E_g) are shown in Table 3.2. (a) Plot n vs. E_g and (b) Plot also n^4 vs. $1/E_g$. What is your conclusion? According to **Moss's rule**, roughly, $n^4 E_g \approx K$, a constant. What is the value of K?

TABLE 3.2 The refractive index n and the bandgap E_g of diamond, Ge, and Si, all of which have the same crystal structure

Material →	Diamond	Silicon	Germanium
Bandgap, E_g (eV)	5	1.1	0.66
n	2.4	3.46	4.0

3.4 Electrons in the CB of a nondegenerate semiconductor

(a) Consider the energy distribution of electrons $n_E(E)$ in the conduction band. Assuming that the density of state $g_{CB}(E) \propto (E - E_c)^{1/2}$ and using Boltzmann statistics $f(E) \approx \exp\left[-(E - E_F)/k_BT\right]$, show that the energy distribution of the electrons in the CB can be written as

$$n_x(x) = Cx^{1/2}\exp(-x)$$

where $x = (E - E_c)/k_BT$ is the electron energy in terms of k_BT measured from E_c and C is a constant at a given temperature (independent of E).

(b) Setting arbitrarily $C = 1$, plot $n_x(x)$ vs. x. Where is the maximum and what is the FWHM (full width at half maximum, that is, between half maximum points)? Is the use of $1.8k_BT$ for the half-maximum width correct?

(c) Show that the average electron energy in the CB is $(3/2)k_BT$, by using the definition of average,

$$x_{\text{average}} = \int_0^\infty x n_x dx \Big/ \int_0^\infty n_x dx$$

where the integration is from $x = 0$ (E_c) to say $x = 10$ (far away from E_c where $n_x \to 0$). You need to use a numerical integration.

(d) Show that the maximum in the energy distribution is at $x = (1/2)$ or at $E_{\max} = (1/2)k_BT$.

3.5 Intrinsic and doped GaAs The properties of GaAs are shown in Table 3.1. Calculate the intrinsic concentration and the intrinsic resistivity at room temperature (take as 300 K). Where is the Fermi level? Assuming the N_c and N_v scale as $T^{3/2}$, what would be the intrinsic concentration at 100°C? If this GaAs crystal is doped with 10^{17} donors cm^{-3} (such as Te), where is the new Fermi level and what is the resistivity of the sample? The drift motilities in GaAs are shown in Table 3.3.

TABLE 3.3 Ionized dopant impurities scatter carriers and reduce the drift mobility. The dependence of μ_e for electrons and μ_h for holes on the total ionized dopant concentration

Dopant concentration (cm^{-3})	0	10^{14}	10^{15}	10^{16}	10^{17}	10^{18}
GaAs, μ_e ($cm^2\ V^{-1}\ s^{-1}$)	8500	–	8000	7000	5000	2400
GaAs, μ_h ($cm^2\ V^{-1}\ s^{-1}$)	400	–	380	310	250	160
Si, μ_e ($cm^2\ V^{-1}\ s^{-1}$)	1450	1420	1370	1200	730	280
Si, μ_h ($cm^2\ V^{-1}\ s^{-1}$)	490	485	478	444	328	157

3.6 Electrons in GaAs Given that the electron effective mass m_e^* for the GaAs is $0.067m_e$, calculate the thermal velocity of the electrons in the CB of a nondegenerately doped GaAs at room temperature (300 K). If μ_e is the drift mobility of the electrons and τ_e the mean free time between electron scattering events (between electrons and lattice vibrations) and if $\mu_e = e\tau_e/m_e^*$, calculate τ_e, given $\mu_e = 8500\,cm^2\ V^{-1}\ s^{-1}$. Calculate the **drift velocity** $v_d = \mu_e E$ of the CB electrons in an applied field E of $10^5\ V\ m^{-1}$. What is your conclusion?

3.7 Extrinsic *n*-GaAs An *n*-type GaAs crystal is doped with 10^{16} donors cm^{-3} (such as Te), what are the electron and hole concentrations, and the conductivity? (See Table 3.3.)

3.8 Extrinsic *n*-Si A Si crystal has been doped *n*-type with $1 \times 10^{17}\,cm^{-3}$ phosphorus (P) donors. The electron drift mobility μ_e depends on the total concentration of ionized dopants N_{dopant}, as in Table 3.3, inasmuch as these ionized dopants scatter the electrons and thereby decrease their drift mobility. What is the conductivity of the sample? Where is the Fermi level with respect to the intrinsic crystal?

3.9 Compensation doping in *n*-type Si An *n*-type Si sample has been doped with 10^{16} phosphorus (P) atoms cm^{-3}. (a) What are the electron and hole concentrations? (b) Calculate the room temperature conductivity of the sample. (c) Where is the Fermi level with respect to E_{Fi}? (d) If we now dope the crystal with 10^{17} boron acceptors, what will be the electron and hole concentrations? (e) Where is the Fermi level with respect to E_{Fi}?

3.10 Free carrier absorption in semiconductors The attenuation of light due to the optical field drifting the free carriers is called **free carrier absorption**. As the free electrons in a semiconductor crystal are accelerated by the optical field, they eventually become scattered by lattice vibrations or impurities, and pass the energy absorbed from the radiation to lattice vibrations. In such cases, ε_r'' and the AC conductivity σ at the same frequency are related by

$$\varepsilon_r'' = \sigma/\varepsilon_o\omega \tag{P3.1}$$

We consider a semiconductor in which the free carriers are electrons (an *n*-type semiconductor). The AC conductivity σ in general is given by

$$\sigma = \sigma_o/(1 + j\omega\tau_e) \tag{P3.2}$$

where σ_o is the DC conductivity, ω is the angular frequency of light, and τ_e is the scattering time of the conduction electrons. σ decreases with frequency in Eq. (P3.2). Consider an *n*-type semiconductor. The free carriers are the electrons in the CB. If the drift mobility of the electrons is μ_e, then $\mu_e = e\tau_e/m_e^*$, where m_e^* is the effective mass of the electrons in the CB of the semiconductor.[28] The DC conductivity $\sigma_o = en\mu_e$, where n is the concentration of CB electrons. Show that the absorption coefficient due to free carrier absorption (due to the conductivity) when $\omega > 1/\tau$ is given by

$$\alpha = \left(\frac{\sigma_o}{nc\varepsilon_o\tau^2}\right)\left(\frac{1}{\omega^2}\right) = \left(\frac{e^3 n}{4\pi^2 nc^3\varepsilon_o m_e^{*2}\mu_e}\right)\lambda^2 \tag{P3.3}$$

[28] As we know, conduction in an *n*-type semiconductor occurs by the drift of free electrons inside the semiconductor crystal. The dopants donate electrons to the crystal, which are free within the crystal.

where n is the refractive index. What would you expect if you plotted α vs. λ^2? Consider a Si crystal doped with 10^{15} cm^{-3} donors. Estimate the free carrier absorption (in m^{-1} and dB m^{-1}) at 1.55 and at 5 μm. What is your conclusion? [Although Eq. (P3.3) is a highly simplified approximation to describing free carrier absorption, it nonetheless provides a rough estimate of its magnitude.]

3.11 GaAs *pn* junction Consider a GaAs *pn* junction that has the following properties: $N_a = 10^{16}$ cm^{-3} (p-side); $N_d = 10^{18}$ cm^{-3} (n-side); $B = 2.0 \times 10^{-16}$ m^3 s^{-1}; cross-sectional area $A = 1.5$ mm $\times$ 1.5 mm. Assume a long diode. What is the diode current due to diffusion in the neutral regions and recombination in the SCL at 300 K when the forward voltage across the diode is 0.8 V and then 1.1 V? (Use the drift motilities in Table 3.3 for calculating the diffusion coefficients through the Einstein relation.)

3.12 InP *pn* junction Consider an InP *pn* junction that has the following properties: $N_a = 10^{15}$ cm^{-3} (p-side); $N_d = 10^{17}$ cm^{-3} (n-side); using $B \approx 4 \times 10^{-16}$ m^3 s^{-1}; cross-sectional area $A = 1$ mm $\times$ 1 mm. Assume a long diode. What is the diode current due to diffusion in the neutral regions and recombination in the SCL at 300 K when the forward voltage across the diode is 0.70 and 0.9 V? The electron mobility in the p-side is about ~6000 cm^2 s^{-1} and the hole mobility on the n-side is roughly ~100 cm^2 s^{-1}. (See also Table 3.1 for n_i and ε_r.) Comment on the ideality factor of this InP *pn* junction.

3.13 Si *pn* junction Consider a long *pn* junction diode with an acceptor doping N_a of 10^{18} cm^{-3} on the p-side and donor concentration of N_d on the n-side. The diode is forward biased and has a voltage of 0.6 V across it. The diode cross-sectional area is 1 mm^2. The minority carrier recombination time, τ, depends on the total dopant concentration, N_{dopant}(cm^{-3}), through the following approximate empirical relation

$$\tau \approx (5 \times 10^{-7})/(1 + 2 \times 10^{-17} N_{\text{dopant}})$$

where τ is in seconds.

(a) Suppose that $N_d = 10^{15}$ cm^{-3}. Then the depletion layer extends essentially into the n-side and we have to consider minority carrier recombination time, τ_h, in this region. Calculate the diffusion and recombination contributions to the total diode current given $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{15}$ cm^{-3}. Use Table 3.3 for μ_e and μ_h. What is your conclusion?

(b) Suppose that $N_d = N_a$. Then W extends equally to both sides and, further, $\tau_e = \tau_h$. Calculate the diffusion and recombination contributions to the diode current given $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{18}$ cm^{-3}. Use Table 3.3 for μ_e and μ_h. What is your conclusion?

3.14 Injected minority carrier charge Consider a *pn* junction with heavier doping on the p-side. The injected minority carriers (holes) represent an *injected excess minority carrier charge* Q_h in the neutral region as shown in Figure 3.24 (a). (There is also excess majority carrier charge so the region is neutral.) Show that

$$Q = I\tau_h \text{ for a long diode and } Q = I\tau_t \text{ for a short diode}$$

in which τ_h is the hole lifetime and τ_t is the diffusion time, or the transit time of holes across the width of the neutral n-region, that is, $\tau_t = l_n^2/2D_h$. What is your conclusion?

3.15 High injection condition The Shockley equation for a *pn* junction under forward bias, as shown in Figure 3.16 (a), was derived by assuming low (weak) injection conditions, that is $p_n(0) \approx \Delta p_n(0) \ll n_{no}$ or N_d on the n-side. Show that when the injection is no longer weak, that is when $p_n(0) \approx n_{no} = N_d$, the applied voltage V reaches V_{SI} (strong injection) given by

$$V_{\text{SI}} = V_o - V_{\text{th}} \ln(N_a/N_d)$$

where V_{th} is the thermal voltage (k_BT/e). Calculate V_o and V_{SI} for a Si *pn* junction that has $N_a = 10^{18}$ cm^{-3} and $N_d = 10^{16}$ cm^{-3}. Can you use the Shockley equation when $V > V_{\text{SI}}$? What happens when $N_a = N_d$? What is your conclusion?

3.16 Heterostructure Consider a Type I heterostructure as shown in Figure 3.27.

(a) If $E_{g1} < E_{g2}$ and if χ_1 and χ_2 are the electron affinities of each semicondcutor, show that

$$\Delta E_c = \chi_1 - \chi_2 \quad \text{and} \quad \Delta E_v = E_{g2} - E_{g1} - \Delta E_c$$

(b) Using the data in Table 3.1, draw the energy band diagram of an *nP* junction between an n-type Ge and P-GaAs. Under forward bias, is it easier to inject electrons or holes?

(c) Draw the energy band diagram for a *pN* junction between p-type Ge and N-GaAs. Under forward bias, is it easier to inject electrons or holes?

3.17 Heterojunction *I–V* characteristics We use some of the data reported by Womac and Rediker (*J. Appl. Phys.*, *43*, 4130, 1972) for AlGaAs/GaAs *pn* heterojunction at 298 K (25°C). Sample A is an N^+p and sample B is a P^+n junction. The *I–V* data of interest are listed in Table 3.4. By a suitable plot find the ideality factor for each. What is your conclusion?

TABLE 3.4 *I–V* **data on two heterojunctions. First set is N^+p and the second set is a P^+n junction**

A: Heterojunction N^+p														
V	0.206 V	0.244	0.290	0.322	0.362	0.412	0.453	0.485	0.537	0.576	0.612	0.662	0.708	
I	1.03 nA	2.07	5.20	10.3	20.7	52.8	105	192	515	1.02 μA	2.03	4.89	10.1	
B: Heterojunction P^+n														
V	0.310 V	0.364	0.402	0.433	0.485	0.521	0.561	0.608	0.682	0.726	0.764	0.807	0.859	0.885
I	2.01 nA	4.91	9.79	19.0	49.5	96.1	194	466	1.96 μ A	5.02	9.75	19.5	50.6	99.5

3.18 AlGaAs LED emitter An AlGaAs LED emitter for in a local optical fiber network has the output spectrum shown in Figure 3.32 (b). It is designed for peak emission at about 822 nm at 25°C. (a) Why does the peak emission wavelength increase with temperature? (b) What is the bandgap of AlGaAs in this LED? (c) The bandgap, E_g, of the ternary $Al_xGa_{1-x}As$ alloys follows the empirical expression, $E_g(eV) = 1.424 + 1.266x + 0.266x^2$. What is the composition of the $Al_xGa_{1-x}As$ in this LED?

3.19 III–V compound semiconductors in optoelectronics Figure 3.52 represents the bandgap E_g and the lattice parameter a in a quaternary III–V alloy system. A line joining two points represents the changes in E_g and a with composition in a ternary alloy composed of the compounds at the ends of that line. For example, starting at GaAs point, $E_g = 1.42$ eV and $a = 0.565$ nm, E_g decreases and a increases as GaAs is alloyed with InAs, as we move

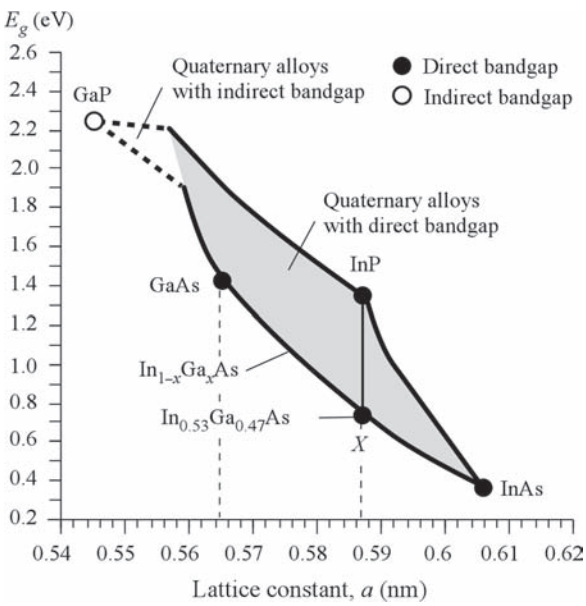
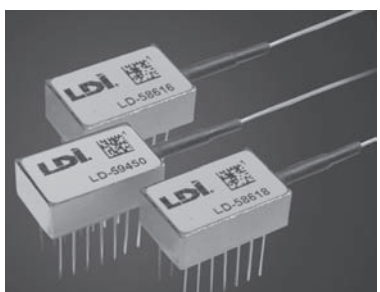


FIGURE 3.52 Bandgap energy E_g and lattice constant a for various III–V alloys of GaP, GaAs, InP, and InAs. A line represents a ternary alloy formed with compounds from the end points of the line. Solid lines are for direct bandgap alloys whereas dashed lines for indirect bandgap alloys. Regions between lines represent quaternary alloys. The line from X to InP represents quaternary alloys $In_{1-x}Ga_xAs_{1-y}P_y$ made from $In_{0.53}Ga_{0.47}As$ and InP, which are lattice-matched to InP.



InGaAsP 1300 nm LED emitters, each pigtailed to an optical fiber for use in optical communication modems and lower speed data/analog transmission systems. *(Courtesy of OSI Laser Diode, Inc.)*

along the line joining GaAs to InAs. Eventually at InAs, $E_g = 0.35\,\text{eV}$ and $a = 0.606\,\text{nm}$. Point X in Figure 3.52 is composed of InAs and GaAs and it is the ternary alloy $In_{1-x}Ga_xAs$. At X, $In_{0.53}Ga_{0.47}As$ (often called "in-gas" in telecom) has $E_g = 0.73\,\text{eV}$ and $a = 0.587\,\text{nm}$, which is the same a as that for InP. $In_{1-x}Ga_xAs$ at X is therefore lattice-matched to InP and can hence be grown on an InP substrate without creating defects at the interface.

Further, $In_{1-x}Ga_x$ As at X can be alloyed with InP to obtain a quaternary alloy[29] $In_{1-x}Ga_xAs_{1-y}P_y$ whose properties lie on the line joining X and InP and therefore all have the same lattice parameter as InP but different bandgap. Layers of $In_{1-x}Ga_xAs_{1-y}P_y$ with composition between X and InP can be grown epitaxially on an InP substrate by various techniques such as liquid phase epitaxy (LPE) or molecular beam epitaxy (MBE). The grey shaded area between the solid lines represents the possible values of E_g and a for the quaternary III–V alloy system in which the bandgap is direct and hence suitable for direct recombination. The compositions of the quaternary alloy lattice matched to InP follow the line from X to InP.

(a) Given that the $In_{1-x}Ga_x$ As at X is $In_{0.53}Ga_{0.47}As$ show that quaternary alloys $In_{1-x}Ga_xAs_{1-y}P_y$ are lattice matched to InP when $y = 1 - 2.13x$.

(b) The bandgap energy E_g, in eV for $In_{1-x}Ga_xAs_{1-y}P_y$ lattice-matched to InP is given by the empirical relation, $E_g(\text{eV}) = 0.75 + 0.46y + 0.14y^2$. Find the composition of the quaternary alloy suitable for an LED emitter operating at 1.30 μm.

3.20 Varshni equation and the change in the bandgap with temperature The Varshni equation describes the change in the energy bandgap E_g of a semiconductor with temperature T as given by Eq. (3.11.2) that is

$$E_g = E_{go} - AT^2/(B + T)$$

where E_{go} is E_g at 0 K, and A and B are constants. Show that

$$\frac{dE_g}{dT} = -\frac{AT(T + 2B)}{(B + T)^2} = -\frac{(E_{go} - E_g)}{T}\left(\frac{T + 2B}{T + B}\right)$$

For GaAs, $E_{go} = 1.519\,\text{eV}$, $A = 5.41 \times 10^{-4}\,\text{eV K}^{-1}$, $B = 204$ K. What is dE_g/dT for GaAs? Find the shift in the emitted wavelength from a GaAs LED per 1°C change at room temperature (300 K). Find the emission wavelength at 27°C and −30°C.

3.21 Dependence on the emission peak and linewidth on temperature Using the Varshni equation find the peak emission wavelength and the linewidth of the emission spectrum from an $In_{0.47}Ga_{0.53}As$ LED when it is cooled from 25°C to −25°C. You can use Eqs. (3.3.1) and (3.3.2). The Varshni constants for $In_{0.47}Ga_{0.53}As$ are $E_{go} = 0.814\,\text{eV}$, $A = 4.906 \times 10^{-4}\,\text{eV K}^{-1}$, $B = 301$ K.

3.22 LED Output Spectrum Given that the width of the relative light intensity vs. photon energy spectrum of an LED is typically around ~3 k_BT, calculate the spectral width in wavelength (nm) of LED emitters operating at 850 nm, 1310 nm, and 1550 nm?

3.23 Linewidth of LEDs Experiments carried out on various direct bandgap semiconductor LEDs give the output spectral linewidth (between half intensity points as listed in Table 3.5). From Figure 3.31 we know that a spread in the wavelength is related to a spread in the photon energy, $\Delta\lambda \approx (hc/E_{\text{ph}}^2)\Delta E_{\text{ph}}$ where $E_{\text{ph}} = h\nu$ is the photon energy. Suppose that we write $E_{\text{ph}} = hc/\lambda$ and $\Delta E_{\text{ph}} = \Delta(h\nu) \approx mk_BT$ where m is a numerical constant. Therefore,

$$\Delta\lambda \approx (mk_BT/hc)\lambda^2 \qquad \textbf{(P3.4)}$$

By appropriately plotting the data in Table 3.5, and assuming $T = 300\,\text{K}$, find m.

TABLE 3.5 Linewidth $\Delta\lambda_{1/2}$ between half points in the output spectrum (spectral intensity vs. wavelength) of eight LEDs using direct bandgap semiconductors

Material (Direct E_g)	AlGaAs	AlGaAs	AlGaAs	GaAs	GaAs	InGaAsP	InGaAsP	InGaAsP
Peak wavelength of emission (λ) nm	650	810	820	890	950	1150	1270	1500
$\Delta\lambda_{1/2}$ nm	22	36	40	50	55	90	110	150

[29] Some books have other formats for the chemical composition for example, $Ga_xIn_{1-x}As_yP_{1-y}$. The present notation $In_{1-x}Ga_xAs_{1-y}P_y$ was chosen to reflect the common vernacular for InGaAs (pronounced "in-gas").

Table 3.6 gives the linewidth $\Delta\lambda_{1/2}$ for various visible LEDs based on GaAsP. Radiative recombination is obtained by appropriately doping the material. Using $m = 3.0$ in Eq. (P3.4), $T = 300$ K, calculate the expected spectral width for each and compare with the experimental value. What is your conclusion?

TABLE 3.6 Linewidth $\Delta\lambda_{1/2}$ between half points in the output spectrum (intensity vs. wavelength) of various visible LEDs using GaAsP

Peak wavelength of emission (λ) nm	565	583	600	635
$\Delta\lambda_{1/2}$ nm	28	36	40	40
Color	Green	Yellow	Orange	Red
Material	GaP(N)	GaAsP(N)	GaAs (N)	GaAsP

3.24 Quantum wells We will consider a quantum well in a semiconductor heterostructure, and assume, for simplicity, an infinite potential energy quantum well with dimensions d along x, and D_y and D_z along y and z directions. The energy of an electron with respect to the bottom of the well is then given by

$$\Delta E = \frac{h^2 n^2}{8m_e^* d^2} + \frac{h^2 n_y^2}{8m_e^* D_y^2} + \frac{h^2 n_z^2}{8m_e^* D_z^2} \qquad \textbf{(P3.5)}$$

Energy in a quantum well

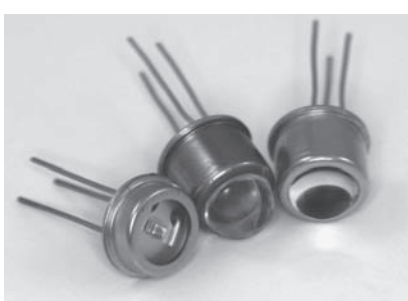
where n, n_y, and n_z are quantum numbers having the values 1, 2, 3, … Consider a well that has $d = 10$ nm, $D_x = D_y = 2\,\mu$m, and $m_e^* = 0.067m_e$. (a) Calculate the minimum energy. Which is the most significant contributing term in Eq. (P3.5) to the lowest energy? (b) What would n_y need to be to get the same energy as the first term? (c) What is the separation δE between the energy levels for motion in the y and z plane? What is your conclusion?

3.25 Energy levels in a quantum well Consider a GaAs QW sandwiched between two $Al_{0.40}Ga_{0.60}As$ layers. The barrier height ΔE_c is approximately 0.30 eV, the electron effective mass in the well is $0.067m_e$, and the width of the QW (d) is 8 nm. Calculate the energy levels ΔE_1 and ΔE_2 from the bottom of E_c, assuming an infinite PE well. Compare these with the calculations for a finite PE well that give $\Delta E_1 = 0.050$ eV and $\Delta E_2 = 0.197$ eV. What is your conclusion?

3.26 QW LED Linewidth Consider the QW-based InGaN (green to blue) and AlGaN (UV) LEDs whose peak emission wavelengths (λ_o) and linewidths (between half intensity points, $\Delta\lambda$) are listed in Table 3.7. For each LED, apply the linewidth equation $\Delta\lambda \approx (mk_BT/hc)\lambda_o^2$ at 300 K, and find m. Do you think this formula can be used for QW LEDs?

TABLE 3.7 Avago ASMT 1W series of InGaN green, cyan, blue, and royal blue LEDs and SETi AlGaN UV LED

LED	Green	Cyan	Blue	Royal blue	UV
λ_o	519	497	454	450	270
$\Delta\lambda$	31.25	29	22.5	20	12



UV emitting LEDs *(Courtesy of Sensor Electronic)*

3.27 External conversion efficiency The total light output power, called the radiant flux, from a number of commercial LEDs is listed in Table 3.8. The table also lists the emitted peak wavelength and the measured linewidth for each LED. For each device calculate the external quantum efficiency (%), EQE, and m in $\Delta\lambda \approx (mk_BT/hc)\lambda^2$. What is your conclusion?

TABLE 3.8 Characteristics of selected commercial LEDs covering the IR, blue, and UV

	IR, Hamamatsu L10823 (InGaAs)	IR, Hamamatsu L10822 (InGaAs)	IR, Osram SFH406 (GaAs)	IR, Vishay TSAL6100 (GaAlAs)	IR, Hamamatsu L1909 (AlGaAs)	Super Blue, Cree (InGaN)	UV, UVTOP 315, SETi (InGaN)
V_F (V)	0.9	1.1	1.25	1.35	1.4	3.8	5.5
I_F (mA)	50	50	40	100	50	30	20
P_o (mW)	2	3.1	7	35	10	1.15	0.6
λ_o (nm)	1650	1300	950	940	870	466	315
$\Delta\lambda$ (nm)	120	90	55	50	80	65	11

3.28 Light extraction efficiency of LEDs Extraction efficiency or ratio depends on the LED structure and the means that have been used to manipulate light to escape from the chip.

(a) Figure 3.53 (a) shows a bare LED chip in which the light from the active region is incident on the semiconductor–air interface. The refractive indices of the semiconductor and ambient are n_s and n_a respectively. Assume a GaN crystal for the semiconductor with a refractive index of 2.5. For normal incidence, what is the transmittance T from the semiconductor to the air? What is the critical angle θ_c at this interface? The fraction of output light from the active region within the emission cone defined by the critical angle θ_c has been evaluated to be $(1/2)[1 - \cos\theta_c]$. Taking into account T as well, estimate the extraction efficiency (EE) if the ambient is air, $n_a = 1$.

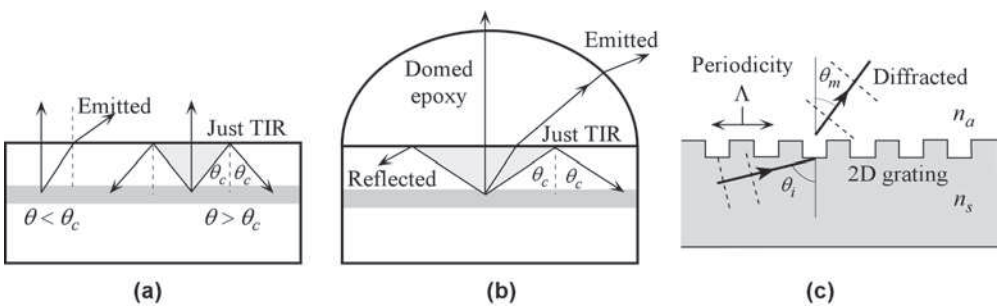


FIGURE 3.53 (a) Total internal reflection results in poor light extraction (b) Improvement in the light extraction by the use of a domed epoxy. (c) A two-dimensional diffraction grating (a photonic crystal) based surface enhances the extraction of light through Bragg diffraction.

(b) Figure 3.53 (b) shows an epoxy dome placed on top of the semiconductor. Suppose that the refractive index of the epoxy is 1.8. Calculate the new critical angle at the semiconductor–epoxy interface and estimate the new EE. What is the critical angle for light rays at the epoxy–air interface? How critical is it to have an exactly hemispherical surface?

(c) Figure 3.53 (c) shows an LED chip whose surface has been patterned to form a two-dimensional (2D) *photonic crystal*, which is a periodic modulation of the refractive index. With the right periodicity Λ, the photonic crystal will leak out the light trapped in the crystal. We will take a highly simplified view and treat the periodic variation as a grating. If θ_i is the angle of incidence, there is a diffracted beam at an angle θ_m, which satisfies the Bragg diffraction condition that involves two different media that is,

$$\Lambda(n_s \sin\theta_i - n_a \sin\theta_m) = m\lambda$$

where λ is the free-space wavelength, m is the diffraction order. Suppose that $n_s \approx 2.5$ (GaN), $n_a = 1$ (air), $\lambda = 500$ nm, $\Lambda = 300$ nm, find the critical angle and the first-order diffracted beams when $\theta_i = 60°$. What is your conclusion?

3.29 LED efficiencies A particular 890 nm IR LED for use in instrumentation has a AlGaAs chip. The active region has been doped p-type with 4×10^{17} cm^{-3} of acceptors and the nonradiative lifetime is about 60 ns. At a forward current of 50 mA, the voltage across it is 1.4 V, and the emitted optical power is 10 mW. Calculate the PCE, IQE, and EQE, and estimate the light extraction ratio. For AlGaAs, $B \approx 1 \times 10^{-16}$ m^3 s^{-1}.

3.30 LED luminous flux

(a) Consider a particular green LED based on InGaN MQW active region. The emission wavelength is 528 nm. At an LED current of 350 mA, the forward voltage is 3.4 V. The emitted luminous flux is 92 lm. Find the power conversion efficiency, external quantum efficiency, luminous efficacy, and the emitted optical power (radiant flux)? (Data for Osram LT CPDP.)

(b) A red LED emits 320 mW of optical power at 656 nm when the current is 400 mA and the forward voltage is 2.15 V. Calculate the power conversion efficiency, external efficiency and the luminous efficacy. (Data for thin film InGaAlP Osram LH W5AM LED.)

(c) A deep blue LED emits at an optical power of 710 mW at 455 nm when the current is 350 mA and the forward voltage is 3.2 V. Calculate the power conversion efficiency, external efficiency, and luminous efficacy. (Data for a GaN Osram LD W5AM LED.)

3.31 SLEDs and ELEDs Experiments carried out on an AlGaAs SLED (surface-emitting LED) and an ELED (edge-emitting LED) give the light output power vs. current data in Table 3.9. (a) Show that the output light power vs. current characteristics are not linear. (b) By plotting the optical power output (P_o) vs. current (I) data on a log-log plot show that $P_o \propto I^n$. Find n for each LED.

TABLE 3.9 Light output power vs. DC current for surface- and edge-emitting LEDs

SLED I (mA)	25	50	75	100	150	200	250	300
SLED Light output power, P_o (mW)	1.04	2.07	3.1	4.06	5.8	7.6	9.0	10.2
ELED I (mA)	25	50	75	100	150	200	250	300
ELED Light output power (mW) P_o	0.46	0.88	1.28	1.66	2.32	2.87	3.39	3.84

3.32 LED-Fiber coupling Efficiency

(a) It is found that approximately 200 μW is coupled into a multimode step index fiber from a surface-emitting LED when the current is 75 mA and the voltage across the LED is about 1.5 V. What is the overall efficiency of operation?

(b) Experiments are carried out on coupling light from a 1310 nm ELED (edge-emitting LED) into multimode and single-mode fibers. (i) At room temperature, when the ELED current is 120 mA, the voltage is 1.3 V and light power coupled into a 50 μm multimode fiber with NA (numerical aperture) of 0.2 is 48 μW. What is the overall efficiency? (ii) At room temperature, when the ELED current is 120 mA, the voltage is 1.3 V and light power coupled into a 9 μm single mode fiber is 7 μW. What is the overall efficiency?

3.33 Internal optical power For a particular AlGaAs LED emitting at 850 nm, the nonradiative recombination lifetime is $\tau_{nr} = 50$ ns. The recombination occurs in the active region, which has been doped with acceptors of concentration 10^{17} cm^{-3}, and the direct recombination coefficient B is 2×10^{-16} cm^3 s^{-1}. What is the internal optical power generated at a current of 100 mA?

3.34 Internal quantum efficiency (IQE) and bandwidth product Show that IQE and the bandwidth f_c product is

$$\eta_{\text{IQE}} f_c = \frac{1}{2\pi\tau_r}$$

What is your conclusion? Consider an LED in which the radiative and nonradiative lifetimes are approximately 10 ns and 50 ns, respectively. What are the IQE and the possible bandwidth f_c? What happens if defects cause τ_{nr} to be 25 ns? What other factors do you think would reduce the bandwidth?

3.35 Bandwidth and rise and fall time

(a) Suppose the rise τ_R for an LED is given, and the rise and fall times are the same. Show that its AC modulation bandwidth f_c (the cutoff frequency) is given by

$$f_c = 0.35/\tau_R$$

(b) The specifications for a particular 1300 nm InGaAs ELED (PE13W series) pigtailed to a fiber has rise and fall times given by 2.5 ns. Calculate the bandwidth f_c. How does that compare with the quoted value 150 MHz?

(c) At large currents, that is, when the minority injection is strong, the radiative recombination time τ_r is no longer constant, but given by Eq. (3.8.7), $\tau_r = 1/B\Delta n$, where Δn is the injected concentration of carriers. The injected carriers (Δn) into the active region are brought in by the current (J) and then they recombine and emit photons, that is, Δn is proportional to the current density J. How would this affect the bandwidth?

3.36 LED electronic drive circuits Figures 3.49 (b) shows one possible way to drive an LED. Assume that V_{CC} is large and that the base-emitter voltage of the BJT is 0.7 V, show that $I_F \approx V/R_F$. How large should V_{CC} be? Consider a red LED that has $V_F = 2\,\text{V}$ at $I_F = 20\,\text{mA}$ and $I_{\max} = 50\,\text{mA}$. If the input control voltage range is from 0 to 5V, what would you recommend for R_F, R_B, and V_{CC}?

3.37 LED electronic drive circuit design Figures 3.54 (a)–(c) shows three possible circuits that can be used to drive a high-current LED from CMOS or similar low-current logic circuits. A bipolar junction transistor (Q_1) is commonly used for this task.

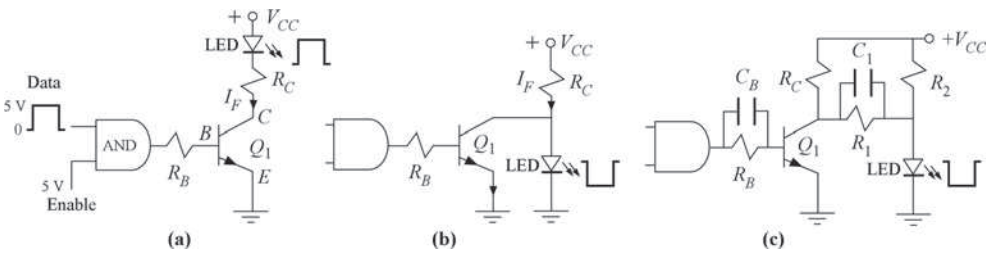
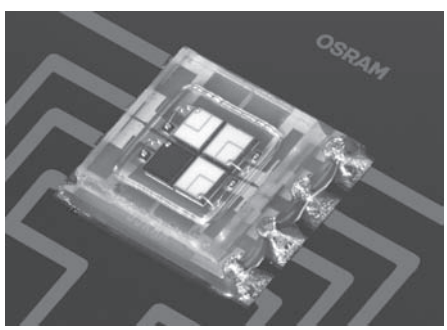


FIGURE 3.54 (a) A CMOS logic circuit (or a gate) drives an LED through a BJT Q_1. (b) The logic gate drives Q_1, which shunts the current from the LED and switches it off. (c) A combination of (a) and (b) with speeding capacitors C_B and C_1 to turn the LED on and off more quickly. In all three cases, the inputs to the logic circuit are the same.

(a) Consider Figure 3.54 (a), if the input is high (5 V), Q_1 is fully turned on and $V_{CE} = 0.25\,\text{V}$ (saturated CE voltage), and $V_{BE} = 0.7\,\text{V}$. What is the current I_F through the LED? If the LED needs $I_F = 50\,\text{mA}$ and $V_F = 2\,\text{V}$, and $V_{CC} = 5\text{V}$, what is R_C? Suppose that Q_1 has a minimum β (or h_{FE}) of 70, what is R_B? Is there an advantage in using higher V_{CC} voltages?

(b) Consider Figure 3.54 (b). Show that the LED is turned off when the input is high (5 V) [as opposed to (a) above]. For the same LED as in (a) with $I_F = 50\,\text{mA}$, $V_F = 2\,\text{V}$, and $V_{CC} = 5\text{V}$, find R_C and R_B.

(c) The circuit Figure 3.54 (c) has been *designed* to drive an LED that has $I_F = 25\,\text{mA}$ and $V_F = 2\,\text{V}$. The resistors are $R_B = 5.6\,\text{k}\Omega$, $R_C = 100\,\Omega$, $R_1 = 22\,\Omega$, and $R_2 = 820\,\Omega$. Explain how the circuit works, assuming first Q_1 is off and then it is on ($V_{CE} = 0.25\,\text{V}$). Is I_F close to 25 mA when Q_1 is off? The capacitors C_B and C_1 across R_B and R_1 speed up the circuit. What do you think C_BR_B and C_1R_1 should be? What are the requirement on Q_1?



This multichip LED from Osram is used in microprojectors. It is based on thin-film GaN technology. *(Courtesy of OSRAM Opto Semiconductors, Germany.)*

A handheld microprojector MPro120 from 3M for projecting photos and videos on a wall uses LEDs for its projection light. *(Courtesy of 3M.)*

LEDs have now widely replaced the incandescent light bulbs in traffic lights.

This LED-based light bulb for use in chandeliers uses a few watts of electric power (2.5W) to generate 135 lumens of light. There are 18 LED chips each with a yellow phosphor to generate the white light.

We thought it [the laser] might have some communications and scientific uses, but we had no application in mind. If we had, it might have hampered us and not worked out as well.

—Arthur Schawlow[1]

Zhores Alferov (on the right) with Valery Kuzmin (technician) in 1971 at the Ioffe Physical Technical Institute, discussing their experiments on heterostructures. Zhores Alferov carried out some of the early pioneering work on heterostructure semiconductor devices that led to the development of a number of important optoelectronic devices, including the heterostructure laser. Zhores Alferov and Herbert Kroemer shared the Nobel Prize in Physics (2000) with Jack Kilby. Their Nobel citation is "for developing semiconductor heterostructures used in high-speed- and opto-electronics." *(Courtesy of Professor Zhores Alferov, Ioffe Physical Technical Institute.)*

[1] Arthur Schawlow (1921–1999; Nobel Laureate, 1981) talking about the invention of the laser.