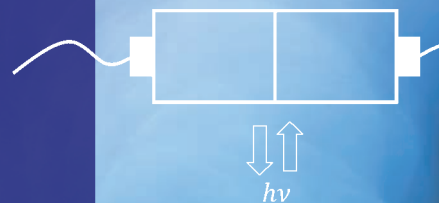# Principles of Solar Cells, LEDs and Diodes

The role of the PN junction

ADRIAN KITAI

WILEY

Principles of Solar Cells, LEDs and Diodes

Principles of Solar Cells, LEDs and Diodes

The role of the PN junction

ADRIAN KITAI

Departments of Engineering Physics and Materials Science and Engineering, McMaster University, Hamilton, Ontario, Canada

Dedicated to my wife Tomoko

Contents

Introduction

Semiconductor devices have revolutionized the way we work and live. Transistors are thought of as one of the most important developments of the twentieth century and they have given rise to the computer age as well as to compact, reliable electronics found in everything from televisions to cell phones.

An even more fundamental semiconductor device exists, however. It is the semiconductor diode or p-n junction diode. Diodes had been developed before the transistor and were used for rectification whereby alternating current can be converted to direct current by employing the unidirectional property of diodes: current normally only flows efficiently in one direction through a diode, and current flow is blocked in the opposite direction. This property of diodes is exploited in power supplies as well as in many other circuits such as those found in radios and limiters. Since an understanding of diodes is required to explain the principles of transistors, diodes are frequently presented as a stepping stone to the transistor.

In the twenty-first century, however, two new major industries are undergoing very rapid developments based directly on the p-n junction diode. Photovoltaic (PV) solar cells and light emitting diodes (LEDs) are both p-n junctions that are designed and optimized to either absorb or emit light. In both cases, an energy conversion process between photons and electrons occurs within a p-n junction.

The consequences of this development constitute a revolution in two major industrial sectors:

1. Energy production has relied on hydrocarbons and nuclear power, and although these will continue to be important, the direct conversion of solar radiation into useful power is the key to a long-term, sustainable energy supply. Ninety-seven percent of all renewable energy on earth is in the form of solar radiation. The beginning of the twenty-first century has seen the abrupt growth of a global solar photovoltaic industry in conjunction with the involvement of governments worldwide, and unprecedented growth in PV production and deployment is now underway. The worldwide consumption of silicon semiconductor material for the entire microelectronics industry is being overtaken by its use for solar cells alone.

2. Electric lighting was achieved by the incandescent lamp in the early twentieth century. In the second half of the twentieth century a significant movement to fluorescent lamps and discharge lamps in which a gas is excited into a plasma that radiates energy more efficiently than a tungsten filament occurred. The twenty-first century will give rise to the virtual displacement of both incandescent lamps and fluorescent lamps by LED lamps. In preparation for this revolution, the world's major lighting companies have

acquired LED research, development and production capabilities. LEDs have already achieved lower power consumption and longer life in small, rugged packages than either incandescent or fluorescent lamps. They also eliminate heavy metals such as mercury used in fluorescent lamps.

The purpose of this book, therefore, is to introduce the physical concepts required for a thorough understanding of p-n junctions starting with semiconductor fundamentals and extending this to the practical implementation of semiconductors in both PV and LED devices. The treatment of a range of important semiconductor materials and device structures is also presented.

The book is aimed at senior undergraduate levels (years three and four). An introductory background in quantum mechanics is assumed, together with general knowledge of junior mathematics, physics and chemistry; however, no background in electronic materials is required. As such this book is designed to be relevant to all engineering students with an interest in semiconductor devices and not specifically to electrical or engineering physics/engineering science students only. This is intentional since solar cells and LEDs involve a wide range of engineering disciplines and should not be regarded as belonging to only one branch of engineering.

In Chapter 1, the physics of solid state electronic materials is covered in detail starting from the basic behaviour of electrons in crystals. The quantitative treatment of electrons and holes in energy bands is presented along with the important concepts of excess carriers that become significant once semiconductor devices are either connected to sources of power or illuminated by light. A series of semiconductor materials and their important properties is also reviewed. The behaviour of semiconductor surfaces and trapping concepts are also introduced since they play an important role in solar cell and LED device performance.

In Chapter 2, the basic physics and important models of a p-n junction device are presented. The approach taken is to present the diode as a semiconductor device that can be understood from the band theory covered in Chapter 1. Various types of diode behaviour, including tunnelling, metal-semiconductor contacts and heterojunctions, are presented as well as reverse breakdown behaviour.

Chapter 3 introduces the theory of photon emission and absorption, a topic that books on semiconductor devices frequently pay less attention to. The standard description that a photon is created when an electron and a hole recombine, or a photon is absorbed when an electron and a hole are generated, is not adequate for a deeper understanding of photon emission and absorption processes. In this chapter the physics of photon creation is explained with a minimum of mathematical complexity, and these concepts are much better understood by following radiation theory and describing the oscillating dipole both classically and using simple quantum mechanics. A section of Chapter 3 describes the exciton relevant to inorganic semiconductors as well as the molecular exciton for organic semiconductors. In addition lineshapes predicted for direct-gap semiconductors are derived. Finally the subject of photometric units introduces the concepts of luminance and colour coordinates that are essential to a discussion of organic and inorganic light emitting diodes.

Chapter 4 covers inorganic solar cells. The concepts regarding the p-n junction introduced in Chapter 2 are further developed to include illumination of the p-n junction and the simplest possible modelling is used to illustrate the behaviour of a solar cell. Then a more realistic solar cell structure and model are presented along with the attendant surface recombination

and absorption issues that must be understood in practical solar cells. A series of solar cell technologies are reviewed starting with bulk single and multicrystalline silicon solar cell technology. Amorphous silicon materials and device concepts are presented. Solar cells made using other semiconductors such as CdTe are introduced followed by multijunction solar cells using layered, lattice-matched III-V semiconductor stacks.

Chapter 5 on inorganic LEDs considers the basic LED structure and its operating principles. The measured lineshape of III-V LEDs is compared with the predictions of Chapter 3. LEDs must be engineered to maximize radiative recombination, and energy loss mechanisms are discussed. The series of developments that marked the evolution of current, high-efficiency LED devices is presented starting from the semiconductors and growth techniques of the 1960s, and following trends in succeeding decades that brought better materials and semiconductor growth methods to the LED industry. The double heterojunction is introduced and the resulting energy well is analysed on the basis of the maximum current density that can be accommodated before it becomes saturated. LED optical outcoupling, which must also be maximized to achieve overall efficiency, is modelled and strategies to optimize outcoupling are discussed. Finally the concept of spectral down-conversion using phosphor materials and the white LED are introduced.

Chapter 6 introduces new concepts required for an understanding of organic semiconductors in general, in which conjugated molecular bonding gives rise to π bands and HOMO and LUMO levels in organic semiconductors. The organic LED is introduced by starting with the simplest single active layer polymer-based LED followed by successively more complex small-molecule LED structures. The roles of the various layers, including electrodes and carrier injection and transport layers, are discussed and the relevant candidate molecular materials are described. Concepts from Chapter 3, including the molecular exciton and singlet and triplet states are used to explain efficiency limitations in the light generation layer of small-molecule OLEDs. In addition the opportunity to use phosphorescent host-guest light emitting layers to improve device efficiency is explained. The organic solar cell is introduced and the concepts of exciton generation and exciton dissociation are described in the context of the heterojunction and the bulk heterojunction. The interest in the use of fullerenes and other related nanostructured materials is explained for the bulk heterojunction.

All the chapters are followed by problem sets that are designed to facilitate familiarity with the concepts and a better understanding of the topics introduced in the chapter. In many cases the problems are quantitative and require calculations; however, a number of more conceptual problems are presented and are designed to give the reader experience in using the Internet or library resources to look up further information. These problems are of particular relevance in Chapters 4, 5 and 6, in which developments in solar cells and LEDs are best understood by referring to the recent literature once the basic concepts are understood.

Adrian Kitai

Acknowledgements

1

Semiconductor Physics

Objectives

1. Understand semiconductor band theory and its relevance to semiconductor devices.
2. Obtain a qualitative understanding of how bands depend on semiconductor materials.
3. Introduce the concept of the Fermi energy.
4. Introduce the concept of the mobile hole in semiconductors.
5. Derive the number of mobile electrons and holes in semiconductor bands.
6. Obtain expressions for the conductivity of semiconductor material based on the electron and hole concentrations and mobilities.
7. Introduce the concepts of doped semiconductors and the resulting electrical characteristics.
8. Understand the concept of excess, non-equilibrium carriers generated by either illumination or by current flow due to an external power supply.
9. Introduce the physics of traps and carrier recombination and generation.
10. Introduce alloy semiconductors and the distinction between direct gap and indirect gap semiconductors.

1.1   Introduction

A fundamental understanding of electron behaviour in crystalline solids is available using the *band theory of solids.* This theory explains a number of fundamental attributes of electrons in solids including:

  (i) concentrations of charge carriers in semiconductors;
 (ii) electrical conductivity in metals and semiconductors;
(iii) optical properties such as absorption and photoluminescence;
(iv) properties associated with junctions and surfaces of semiconductors and metals.

    The aim of this chapter is to present the theory of the band model, and then to exploit it to describe the important electronic properties of semiconductors. This is essential for a proper understanding of p-n junction devices, which constitute both the photovoltaic (PV) solar cell and the light-emitting diode (LED).

1.2   The Band Theory of Solids

There are several ways of explaining the existence of energy bands in crystalline solids. The simplest picture is to consider a single atom with its set of discrete energy levels for its electrons. The electrons occupy quantum states with quantum numbers n, l, m and s denoting the energy level, orbital and spin state of the electrons. Now if a number N of identical atoms are brought together in very close proximity as in a crystal, there is some degree of spatial overlap of the outer electron orbitals. This means that there is a chance that

any pair of these outer electrons from adjacent atoms could trade places. The Pauli exclusion principle, however, requires that each electron occupy a unique energy state. Satisfying the Pauli exclusion principle becomes an issue because electrons that trade places effectively occupy new, *spatially extended* energy states. The two electrons apparently occupy the same spatially extended energy state.

In fact, since outer electrons from all adjacent atoms may trade places, outer electrons from *all* the atoms may effectively trade places with each other and therefore a set of outermost electrons from the *N* atoms all appear to share a spatially extended energy state that extends through the entire crystal. The Pauli exclusion principle can only be satisfied if these electrons occupy a set of *distinct*, spatially extended energy states. This leads to a set of slightly different energy levels for the electrons that all originated from the same atomic orbital. We say that the atomic orbital splits into an *energy band* containing a set of electron states having a set of closely spaced energy levels. Additional energy bands will exist if there is some degree of spatial overlap of the atomic electrons in lower-lying atomic orbitals. This results in a set of energy bands in the crystal. Electrons in the lowest-lying atomic orbitals will remain virtually unaltered since there is virtually no spatial overlap of these electrons in the crystal.

The picture we have presented is conceptually a very useful one and it suggests that electrical conductivity may arise in a crystal due to the formation of spatially extended electron states. It does not directly allow us to quantify and understand important details of the behaviour of these electrons, however.

We need to understand the behaviour in a solid of the electrons that move about in the material. These mobile charge carriers are crucially important in terms of the electrical properties of devices. An electron inside an infinitely large vacuum chamber is a free electron, but a mobile electron in a solid behaves very differently.

We can obtain a more detailed model as follows. The mobile electrons in a crystalline semiconductor are influenced by the electric potential in the material. This potential has a spatial periodicity on an atomic scale due to the crystal structure. For example, positively charged atomic sites provide potential valleys to a mobile electron and negatively charged atomic sites provide potential peaks or barriers. In addition, the semiconductor is finite in its spatial dimensions and there will be additional potential barriers or potential changes at the boundaries of the semiconductor material.

The quantitative description of these spatially extended electrons requires the use of wavefunctions that include their spatial distribution as well as their energy and momentum. These wavefunctions may be obtained by solving Schrödinger's equation. The following section presents a very useful band theory of crystalline solids and the results.

1.3   The Kronig–Penney Model

The *Kronig–Penney model* is able to explain the essential features of band theory.

First, consider an electron that can travel within a one-dimensional periodic potential $V(x)$. The periodic potential can be considered as a series of regions having zero potential energy separated by potential energy barriers of height V_0, as shown in Figure 1.1, forming a simple periodic potential with period $a + b$. We associate $a + b$ also with the lattice constant of the crystal. Note that the electric potential in a real crystal does not exhibit the
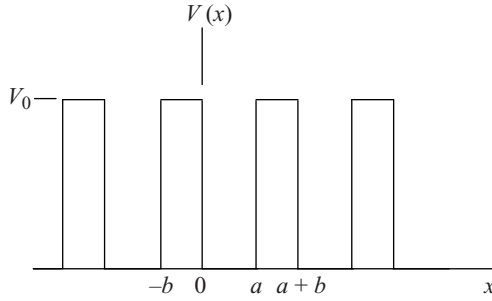
Figure 1.1   *Simple one-dimensional potential V(x) used in the Kronig–Penney model*

idealized shape of this periodic potential; however, the result turns out to be relevant in any case, and Schrödinger's equation is much easier to solve starting from the potential of Figure 1.1.

In order to obtain the electron wavefunctions relevant to an electron in the crystalline solid, $V(x)$ is substituted into the time-independent form of Schrödinger's equation:

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x) = E\psi(x) \tag{1.1}$$

where $V(x)$ is the potential energy and E is total energy.

For $0 \leq x \leq a$ we have $V = 0$ and the general solution to Equation 1.1 yields:

$$\psi(x) = Ae^{iKx} + Be^{-iKx} \tag{1.2a}$$

where

$$E = \frac{\hbar^2 K^2}{2m} \tag{1.2b}$$

For $-b \leq x \leq 0$ we have

$$\psi(x) = Ce^{Qx} + De^{-Qx} \tag{1.3a}$$

where

$$V_0 - E = \frac{\hbar^2 Q^2}{2m} \tag{1.3b}$$

Boundary conditions must be satisfied such that $\psi(x)$ and $\frac{d\psi(x)}{dx}$ are continuous functions. At $x = 0$, equating (1.2a) and (1.3a), we have

$$A + B = C + D \tag{1.4a}$$

and equating derivatives of (1.2a) and (1.3a),

$$iK(A - B) = Q(C - D) \tag{1.4b}$$

An important additional constraint on the required wavefunctions results from the periodicity of the lattice. The solution to Equation 1.1 for any periodic potential must also have the form of a *Bloch function*:

$$\psi(x) = u_k(x)e^{ikx} \tag{1.5}$$

Here, k is the wavenumber of a plane wave. There are no restrictions on this wavenumber; however, $u_k(x)$ must be a periodic function with the same periodicity as the lattice.

Consider two x-values separated by one lattice constant, namely $x = -b$ and $x = a$. Now, Equation 1.5 states that $\psi(x + a + b) = \psi(x)e^{ik(a+b)}$. At $x = -b$ this may be written as:

$$\psi(a) = \psi(-b)e^{ik(a+b)} \tag{1.6}$$

The boundary conditions to satisfy $\psi(x)$ and $\frac{d\psi(x)}{dx}$ being continuous functions at $x = a$ may now be written by substituting ψ from Equations 1.2 and 1.3 into Equation 1.6:

$$Ae^{iKa} + Be^{-iKa} = (Ce^{-Qb} + De^{Qb})e^{ik(a+b)} \tag{1.7a}$$

and substituting the corresponding derivatives:

$$iK(Ae^{iKa} - Be^{-iKa}) = Q(Ce^{-Qb} + De^{Qb})e^{ik(a+b)} \tag{1.7b}$$

Equations 1.4a, 1.4b, 1.7a and 1.7b constitute four equations with four unknowns A, B, C and D. A solution exists only if the determinant of the coefficients of A, B, C and D is zero (Cramer's rule). This requires that

$$\frac{Q^2 - K^2}{2QK} \sinh QB \sin Ka + \cosh Qb \cos Ka = \cos k(a + b) \tag{1.7c}$$

This may be simplified if the limit $b \to 0$ and $V_0 \to \infty$ is taken such that bV_0 is constant (see Problem 1.1). We now define

$$P = \frac{Q^2 ba}{2}$$

Since $Q \gg K$ and $Qb \ll 1$ we obtain

$$\cos ka = P\frac{\sin Ka}{Ka} + \cos Ka \tag{1.8}$$

Here k is the wavevector of the electron describing its momentum $p = \hbar k$ and

$$K = \frac{1}{\hbar}\sqrt{2mE} \tag{1.9}$$

which means that K is a term associated with the electron's energy.

Now, Equation 1.8 only has solutions if the righthand side of Equation 1.8 is between -1 and $+1$, which restricts the possible values of Ka. The righthand side is plotted as a function of Ka in Figure 1.2.

$$P\frac{\sin Ka}{Ka} + \cos Ka$$

Solutions exist   Solutions exist

iii   ii   i   i   ii   iii

+1

Ka

−1

-3π -2π $-\pi$ 0 π 2π 3π
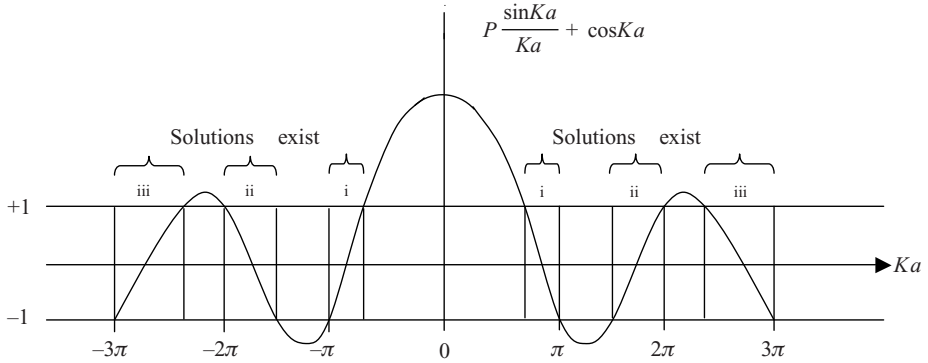
Figure 1.2 *Graph of righthand side of Equation 1.8 as a function of P for P = 2*

Since *K* and *E* are related by Equation 1.9, these allowed ranges of *Ka* actually describe *energy bands* (allowed ranges of *E*) separated by *energy gaps* (forbidden ranges of *E*). *Ka* may be re-plotted on an energy axis, which is related to the *Ka* axis by the square root relationship of Equation 1.9. It is convenient to view *E* on a vertical axis as a variable dependent on *k*. Note that $k = \frac{n\pi}{a}$ for integer values of *n* at the edges of each energy band where the left side of Equation 1.8 is equal to ± 1. These critical values of *k* occur at the boundaries of what are called *Brillouin zones*. A sketch of *E* versus *k* is shown in Figure 1.3, which clearly shows the energy bands and energy gaps.

Let us now plot the free electron graph for *E* versus *k*. Solving Equation 1.1 for a free electron with $V = 0$ yields the solution

$$\psi(x) = Ae^{ikx} + Be^{-ikx}$$

E

k

$-\dfrac{3\pi}{a}$ $-\dfrac{2\pi}{a}$ $-\dfrac{\pi}{a}$ 0 $\dfrac{\pi}{a}$ $\dfrac{2\pi}{a}$ $\dfrac{3\pi}{a}$
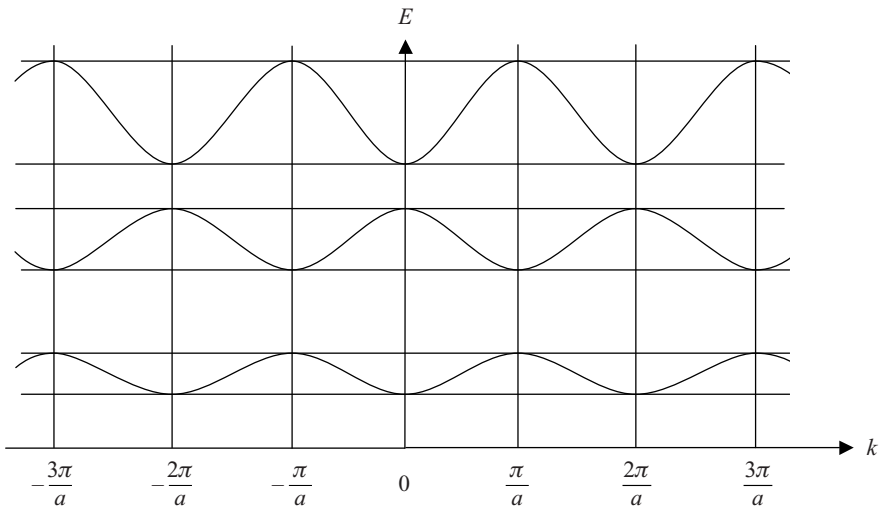
Figure 1.3 *Plot of* E *versus* k *showing how* k *varies within each energy band and the existence of energy bands and energy gaps. The vertical lines at* $k = n\frac{\pi}{a}$ *are Brillouin zone boundaries*
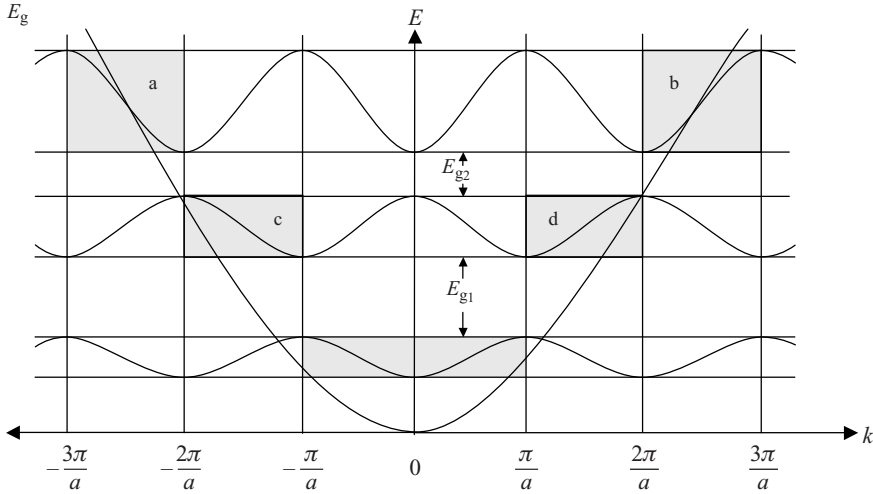
Figure 1.4 *Plot of* E *versus* k *comparing the result of the Kronig–Penney model to the free electron parabolic result*

where

$$E = \frac{\hbar^2 k^2}{2m} \tag{1.10}$$

This parabolic E versus k relationship is plotted superimposed on the curves from Figure 1.3. The result is shown in Figure 1.4.

Taking the limit $P \rightarrow 0$, and combining Equations 1.8 and 1.9, we obtain:

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

which is identical to Equation 1.10. This means that the dependence of E on k in Figure1.4 approaches a parabola as expected if the amplitude of the periodic potential is reduced to zero. In fact, the relationship between the parabola and the Kronig–Penney model is evident if we look at the solutions to Equation 1.4 within the shaded regions in Figure 1.4 and regard them as portions of the parabola that have been broken up by energy gaps and distorted in shape. For a weak periodic potential (small P) the solutions to Equation 1.4 would more closely resemble the parabola. We refer to Equation 1.10 as a *dispersion relation* – it relates energy to the wavenumber of a particle.

At this point, we can draw some very useful conclusions based on the following result: *The size of the energy gaps increases as the periodic potential increases in amplitude in a crystalline solid.* Periodic potentials are larger in amplitude for crystalline semiconductors that have small atoms since there are then fewer atomically bound electrons to screen the point charges of the nuclei of the atoms. In addition, periodic potentials increase in amplitude for compound semiconductors as the ionic character of the crystal bonding increases. This will be illustrated in Section 1.10 for some real semiconductors.

To extend our understanding of energy bands we now need to turn to another picture of electron behaviour in a crystal.

1.4   The Bragg Model

Since electrons behave like waves, they will exhibit the behaviour of waves that undergo reflections. Notice that in a crystal with lattice constant a, the Brillouin zone boundaries occur at

$$k = \frac{n\pi}{a} = \frac{2\pi}{\lambda}$$

which may be rearranged to obtain

$$2a = n\lambda$$

The well-known Bragg condition relevant to waves that undergo strong reflections when incident on a crystal with lattice constant a is

$$2a \sin \theta = n\lambda$$

Now, if the electron is treated as a wave incident at $\theta = 90°$ then we have

$$2a = n\lambda$$

which is precisely the case at Brillouin zone boundaries. We therefore make the following observation: Brillouin zone boundaries occur when the electron wavelength satisfies the requirement for strong reflections from crystal lattice planes according to the Bragg condition. The free electron parabola in Figure 1.4 is similar to the Kronig–Penney model in the shaded regions well away from Brillouin zone boundaries; however, as we approach Brillouin zone boundaries, strong deviations take place and energy gaps are observed.

There is therefore a fundamental connection between the Bragg condition and the formation of energy gaps. The electrons that satisfy the Bragg condition actually exist as *standing waves* since reflections will occur equally for electrons travelling in both directions of the x axis, and standing waves do not travel. Provided electrons have wavelengths not close to the Bragg condition, they interact relatively weakly with the crystal lattice and behave more like free electrons.

The E versus k dependence immediately above and below any particular energy gap is contained in four shaded regions in Figure 1.4. For example, the relevant shaded regions for E_{g2} in Figure 1.4 are labelled a, b, c and d. These four regions are redrawn in Figure 1.5. Energy gap E_{g2} occurs at $k = \pm \frac{2\pi}{a}$. Since this is a standing wave condition with both electron velocity and electron momentum $p = \hbar k$ equal to zero, E_{g2} is redrawn at $k = 0$ in Figure 1.5. Since we are only interested in relative energies, the origin of the energy axis is moved for convenience, and we can arbitrarily redefine the origin of the energy axis. Figure 1.5 is known as a *reduced zone scheme*.

1.5   Effective Mass

We now introduce the concept of *effective mass* m^* to allow us to quantify electron behaviour. Effective mass changes in a peculiar fashion near Brillouin zone boundaries, and generally is not the same as the free electron mass m. It is easy to understand that the effective acceleration of an electron in a crystal due to an applied electric field will depend
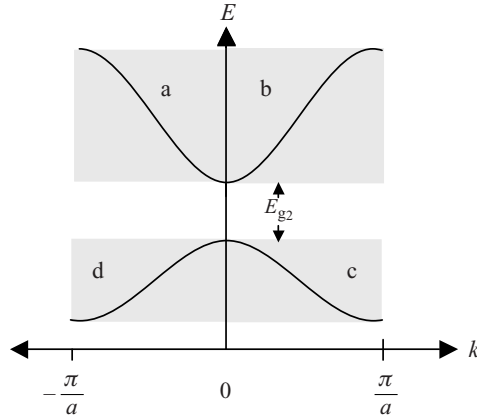
Figure 1.5   *Plot of* E *versus* k *in reduced zone scheme taken from regions a, b, c and d in Figure 1.4*

strongly on the nature of the reflections of electron waves off crystal planes. Rather than trying to calculate the specific reflections for each electron, we instead modify the mass of the electron to account for its observed willingness to accelerate in the presence of an applied force.

To calculate m^* we start with the free electron relationship

$$E = \frac{1}{2}mv_{\mathrm{g}}^2$$

where v_{g} is the group velocity of the electron. Upon differentiation with respect to k,

$$\frac{\mathrm{d}E}{\mathrm{d}k} = mv_{\mathrm{g}}\frac{\mathrm{d}v_{\mathrm{g}}}{\mathrm{d}k} \tag{1.11}$$

Since $p = \hbar k = mv_{\mathrm{g}}$ we can write

$$\frac{v_{\mathrm{g}}}{k} = \frac{\mathrm{d}v_{\mathrm{g}}}{\mathrm{d}k} = \frac{\hbar}{m} \tag{1.12}$$

Combining Equations 1.11 and 1.12 we obtain

$$\frac{\mathrm{d}E}{\mathrm{d}k} = v_{\mathrm{g}}\hbar$$

or

$$v_{\mathrm{g}} = \frac{1}{\hbar}\frac{\mathrm{d}E}{\mathrm{d}k} \tag{1.13}$$

Note that the group velocity falls to zero at the Brillouin zone boundaries where the slope of the E versus k graph is zero. This is consistent with the case of a standing wave.

Now, using Newton's law,

$$F = \frac{\mathrm{d}p}{\mathrm{d}t} = \hbar\frac{\mathrm{d}k}{\mathrm{d}t} \tag{1.14}$$

From Equations 1.13 and 1.14, we can write

$$\frac{dv_g}{dt} = \frac{1}{\hbar}\frac{d^2 E}{dk\,dt} = \frac{1}{\hbar}\frac{d^2 E}{dk^2}\frac{dk}{dt} = \frac{F}{\hbar^2}\frac{d^2 E}{dk^2} \tag{1.15}$$

If we assign m^* to represent an effective electron mass, then Newton's law tells us that

$$\frac{dv_g}{dt} = \frac{F}{m^*}$$

Upon examination Equation 1.15 actually expresses Newton's law provided we define

$$m^* = \frac{\hbar^2}{\frac{d^2 E}{dk^2}} \tag{1.16}$$

Since $\frac{d^2 E}{dk^2}$ is the curvature of the plot in Figure 1.5, it is interesting to note that m^* will be negative for certain values of k. This may be understood physically: if an electron that is close to the Bragg condition is accelerated slightly by an applied force it may then move even closer to the Bragg condition, reflect more strongly off the lattice planes, and effectively accelerate in the direction opposite to the applied force.

We can apply Equation 1.16 to the free electron case where $E = \frac{\hbar^2 k^2}{2m}$ and we immediately see that $m^* = m$ as expected. In addition at the bottom or top of energy bands illustrated in Figure 1.5, the shape of the band may be approximated as parabolic for small values of k and hence a constant effective mass is often sufficient to describe electron behaviour for small values of k. This will be useful when we calculate the number of electrons in an energy band.

1.6   Number of States in a Band

The curves in Figure 1.5 are misleading in that electron states in real crystals are discrete and only a finite number of states exist within each energy band. This means that the curves should be regarded as closely spaced dots that represent quantum states. We can determine the number of states in a band by considering a semiconductor crystal of length L and modelling the crystal as an infinite-walled potential box of length L with a potential of zero inside the well. See Example 1.1.

Example 1.1

An electron is inside a potential box of length L with infinite walls and zero potential in the box. The box is shown below.
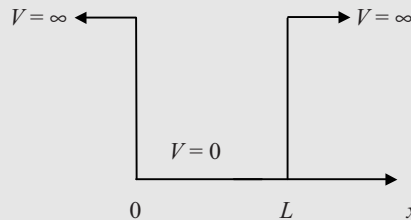
(a) Find the allowed energy levels in the box.
(b) Find the wavefunctions of these electrons.

Solution

(a) Inside the box, from Schrödinger's equation, we can substitute $V(x) = 0$ and we obtain

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi(x)}{\mathrm{d}x^2} = E\psi(x)$$

Solutions are of the form

$$\psi(x) = A \exp\frac{i\sqrt{2mE}}{\hbar}x + B \exp\frac{-i\sqrt{2mE}}{\hbar}x$$

In regions where $V = \infty$ the wavefunction is zero. In order to avoid discontinuities in the wavefunction we satisfy boundary conditions at $x = 0$ and at $x = L$ and require that $\psi(0) = 0$ and $\psi(L) = 0$. These boundary conditions can be written

$$0 = A + B \text{ or } B = -A$$

and

$$0 = A \exp\frac{i\sqrt{2mE}}{\hbar}L + B \exp\frac{-i\sqrt{2mE}}{\hbar}L$$

$$= A\left(\exp\frac{i\sqrt{2mE}}{\hbar}L - \exp\frac{-i\sqrt{2mE}}{\hbar}L\right) = C\sin\frac{\sqrt{2mE}}{\hbar}L$$

where C is a constant. Now $\sin\theta$ is zero provided $\theta = n\pi$ where n is an integer and hence

$$\frac{\sqrt{2mE}}{\hbar}L = n\pi$$

A discrete set of allowed energy values is obtained by solving for E to obtain

$$E_{\mathrm{n}} = \frac{n^2\pi^2\hbar^2}{2mL^2}$$

(b) The corresponding wavefunctions may be found by substituting the allowed energy values into Schrödinger's equation and solving:

$$-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2\psi(x)}{\mathrm{d}x^2} = \frac{n^2\pi^2\hbar^2}{2mL^2}\psi(x)$$

now

$$\frac{\mathrm{d}^2 \psi(x)}{\mathrm{d}x^2} = -\frac{n^2 \pi^2}{L^2} \psi(x)$$

and hence

$$\psi_n(x) = A \sin\left(\frac{n\pi}{L}x\right)$$

From Example 1.1 we obtain

$$\psi_n(x) = A \, \sin\frac{n\pi}{L}x \tag{1.17}$$

where n is a quantum number, and

$$k = \frac{n\pi}{L}, n = 1, 2, 3 \ldots$$

As n increases we will inevitably reach the k value corresponding to the Brillouin zone boundary from the band model

$$k = \frac{\pi}{a}$$

This will occur when

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

and therefore $n = \frac{L}{a}$. The maximum possible value of n now becomes the macroscopic length of the semiconductor crystal divided by the unit cell dimension, which is simply the number of unit cells in the crystal, which we shall call N. Since electrons have an additional quantum number s (spin quantum number) that may be either $\frac{1}{2}$ or $-\frac{1}{2}$, the maximum number of electrons that can occupy an energy band becomes

$$n = 2N$$

Although we have considered a one-dimensional model, the results can readily be extended into two or three dimensions and we still obtain the same result. See Problem 1.3.

We are now ready to determine the actual number of electrons in a band, which will allow us to understand electrical conductivity in semiconductor materials.

1.7   Band Filling

The existence of $2N$ electron states in a band does not determine the actual number of electrons in the band. At low temperatures, the electrons will occupy the lowest allowed energy levels, and in a semiconductor like silicon, which has 14 electrons per atom, several low-lying energy bands will be filled. In addition, the highest occupied energy band will be full, and then the next energy band will be empty. This occurs because silicon has an even number of valence electrons per unit cell, and when there are N unit cells, there will
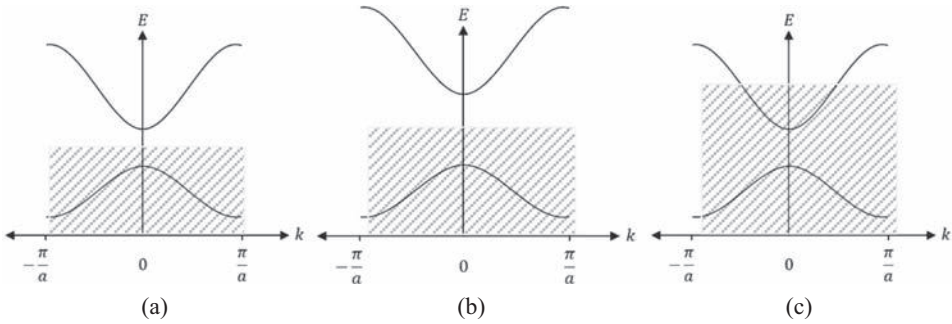
Figure 1.6   *The degree of filling of the energy bands in (a) semiconductors, (b) insulators and (c) metals at temperatures approaching 0 K. Available electron states in the hatched regions are filled with electrons and the energy states at higher energies are empty*

be the correct number of electrons to fill the $2N$ states in the highest occupied energy band. A similar argument occurs for germanium as well as carbon (diamond) although diamond is an insulator due to its large energy gap.

Compound semiconductors such as GaAs and other III-V semiconductors as well as CdS and other II-VI semiconductors exhibit the same result: The total number of electrons per unit cell is even, and at very low temperatures in a semiconductor the highest occupied band is filled and the next higher band is empty.

In many other crystalline solids this is not the case. For example group III elements Al, Ga and In have an odd number of electrons per unit cell, resulting in the highest occupied band being half filled since the $2N$ states in this band will only have N electrons to fill them. These are metals. Figure 1.6 illustrates the cases we have described, showing the electron filling picture in semiconductors, insulators and metals.

In Figure 1.6a the highest filled band is separated from the lowest empty band by an energy gap E_g that is typically in the range from less than 1 eV to between 3 and 4 eV in semiconductors. A completely filled energy band will not result in electrical conductivity because for each electron with positive momentum $p = \hbar k$ there will be one having negative momentum $p = -\hbar k$ resulting in no net electron momentum and hence no net electron flux even if an electric field is applied to the material.

Electrons may be promoted across the energy gap E_g by thermal energy or optical energy, in which case the filled band is no longer completely full and the empty band is no longer completely empty, and now electrical conduction occurs.

Above this range of E_g lie insulators (Figure 1.6b), which typically have an E_g in the range from about 4 eV to over 6 eV. In these materials it is difficult to promote electrons across the energy gap.

In metals, Figure 1.6c shows a partly filled energy band as the highest occupied band. The energy gap has almost no influence on electrical properties whereas occupied and vacant electron states within this partly filled band are significant: strong electron conduction takes place in metals because empty states exist in the highest occupied band, and electrons may be promoted very easily into higher energy states within this band. A very small applied electric field is enough to promote some electrons into higher energy states that impart a

net momentum to the electrons within the band and an electron flow results, which results in the high electrical conductivity in metals.

1.8   Fermi Energy and Holes

Of particular interest is the existence in semiconductors, at moderate temperatures such as room temperature, of the two energy bands that are partly filled. The higher of these two bands is mostly empty but a number of electrons exist near the bottom of the band, and the band is named the *conduction band* because a net electron flux or flow may be obtained in this band. The lower band is almost full; however, because there are empty states near the top of this band, it also exhibits conduction and is named the *valence band*. The electrons that occupy it are valence electrons, which form covalent bonds in a semiconductor such as silicon. Figure 1.7 shows the room temperature picture of a semiconductor in thermal equilibrium. An imaginary horizontal line at energy E_f, called the *Fermi energy*, represents an energy above which the probability of electron states being filled is under 50%, and below which the probability of electron states being filled is over 50%. We call the empty states in the valence band *holes*. Both valence band holes and conduction band electrons contribute to conductivity.

In a semiconductor we can illustrate the valence band using Figure 1.8, which shows a simplified two-dimensional view of silicon atoms bonded covalently. Each covalent bond requires two electrons. The electrons in each bond are not unique to a given bond, and are shared between all the covalent bonds in the crystal, which means that the electron wavefunctions extend spatially throughout the crystal as described in the Kronig–Penney model. A valence electron can be thermally or optically excited and may leave a bond to form an *electron-hole pair* (EHP). The energy required for this is the bandgap energy of the
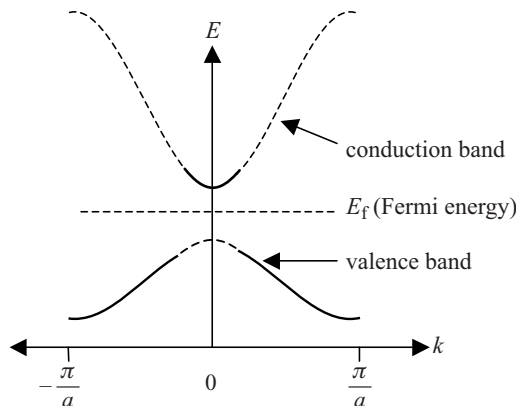
Figure 1.7   *Room temperature semiconductor showing the partial filling of the conduction band and partial emptying of the valence band. Valence band holes are formed due to electrons being promoted across the energy gap. The Fermi energy lies between the bands. Solid lines represent energy states that have a significant chance of being filled*
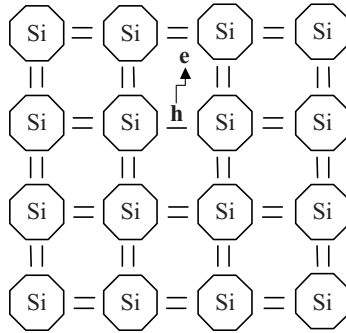
Figure 1.8 *Silicon atoms have four covalent bonds as shown. Although silicon bonds are tetrahedral, they are illustrated in two dimensions for simplicity. Each bond requires two electrons, and an electron may be excited across the energy gap to result in both a hole in the valence band and an electron in the conduction band that are free to move independently of each other*

semiconductor. Once the electron leaves a covalent bond a hole is created. Since valence electrons are shared, the hole is likewise shared among bonds and is able to move through the crystal. At the same time the electron that was excited enters the conduction band and is also able to move through the crystal resulting in two independent charge carriers.

In order to calculate the conductivity arising from a particular energy band, we need to know the number of electrons n per unit volume of semiconductor, and the number of holes p per unit volume of semiconductor resulting from the excitation of electrons across the energy gap E_g. In the special case of a pure or *intrinsic* semiconductor, we can write the carrier concentrations as n_i and p_i such that $n_i = p_i$

1.9   Carrier Concentration

The determination of n and p requires us to find the number of states in the band that have a significant probability of being occupied by an electron, and for each state we need to determine the probability of occupancy to give an appropriate weighting to the state.

We will assume a constant effective mass for the electrons or holes in a given energy band. In real semiconductor materials the relevant band states are either near the top of the valence band or near the bottom of the conduction band as illustrated in Figure 1.7. In both cases the band shape may be approximated by a parabola, which yields a constant curvature and hence a constant effective mass as expressed in Equation 1.16.

In contrast to effective mass, the probability of occupancy by an electron in each energy state depends strongly on energy, and we cannot assume a fixed value. We use the Fermi–Dirac distribution function, which may be derived from Boltzmann statistics as follows. Consider a crystal lattice having lattice vibrations, or *phonons*, that transfer energy to electrons in the crystal. These electrons occupy quantum states that can also transfer energy back to the lattice, and a thermal equilibrium will be established.

Consider an electron in a crystal that may occupy lower and higher energy states E_1^e and E_2^e respectively, and a lattice phonon that may occupy lower and higher energy states E_1^p and E_2^p respectively. Assume this electron makes a transition from energy E_1^e to E_2^e by accepting energy from the lattice phonon while the phonon makes a transition from E_2^p to E_1^p. For conservation of energy,

$$E_2^e - E_1^e = E_2^p - E_1^p \tag{1.18}$$

The probability of these transitions occurring can now be analysed. Let $p(E^e)$ be the probability that the electron occupies a state having energy E^e. Let $p(E^p)$ be the probability that the phonon occupies an energy state having energy E^p. For a system in thermal equilibrium the probability of an electron transition from E_1^e to E_2^e is the same as the probability of a transition from E_2^e to E_1^e, and we can write

$$p(E_2^p)p(E_1^e)(1 - p(E_2^e)) = p(E_1^p)p(E_2^e)(1 - p(E_1^e)) \tag{1.19}$$

because the probability that an electron makes a transition from E_1^e to E_2^e is proportional to the terms on the lefthand side in which the phonon at E_2^p must be available and the electron at E_1^e must be available. In addition, the electron state at E_2^e must be vacant because electrons, unlike phonons, must obey the Pauli exclusion principle, which allows only one electron per quantum state. Similarly the probability that the electron makes a transition from E_2^e to E_1^e is proportional to the terms on the righthand side.

From Boltzmann statistics (see Appendix 3) for phonons or lattice vibrations we use the Boltzmann distribution function:

$$p(E) \propto \exp\left(-\frac{E}{kT}\right) \tag{1.20}$$

Combining Equations 1.19 and 1.20 we obtain

$$\exp\left(-\frac{E_2^p}{kT}\right)p(E_1^e)(1 - p(E_2^e)) = \exp\left(-\frac{E_1^p}{kT}\right)p(E_2^e)(1 - p(E_1^e))$$

which may be written

$$p(E_1^e)(1 - p(E_2^e)) = \exp\left(\frac{E_2^p - E_1^p}{kT}\right)p(E_2^e)(1 - p(E_1^e))$$

Using Equation 1.18 this can be expressed entirely in terms of electron energy levels as

$$p(E_1^e)(1 - p(E_2^e)) = \exp\left(\frac{E_2^e - E_1^e}{kT}\right)p(E_2^e)(1 - p(E_1^e))$$

Rearranging this we obtain

$$\frac{p(E_1^e)}{1 - p(E_1^e)}\exp\left(\frac{E_1^e}{kT}\right) = \frac{p(E_2^e)}{1 - p(E_2^e)}\exp\left(\frac{E_2^e}{kT}\right) \tag{1.21}$$

The left side of this equation is a function only of the initial electron energy level and the right side is only a function of the final electron energy level. Since the equation must always hold and the initial and final energies may be chosen arbitrarily we must conclude that both sides of the equation are equal to an energy-independent quantity, which can only

be a function of the remaining variable T. Let this quantity be $f(T)$. Hence using either the left side or the right side of the equation we can write

$$\frac{p(E)}{1 - p(E)} \exp\left(\frac{E}{kT}\right) = f(T)$$

where E represents the electron energy level.

Solving for $p(E)$ we obtain

$$p(E) = \frac{1}{1 + \frac{1}{f(T)} \exp\left(\frac{E}{kT}\right)} \tag{1.22}$$

We now formally define the Fermi energy E_f to be the energy level at which $p(E) = \frac{1}{2}$ and hence

$$\frac{1}{f(T)} \exp\left(\frac{E_f}{kT}\right) = 1$$

or

$$\frac{1}{f(T)} = \exp\left(\frac{-E_f}{kT}\right)$$

Under equilibrium conditions the final form of the probability of occupancy at temperature T for an electron state having energy E is now obtained by substituting this into Equation 1.22 to obtain

$$F(E) = \frac{1}{1 + \exp\left(\frac{E - E_f}{kT}\right)} \tag{1.23}$$

where $F(E)$ is used in place of $p(E)$ to indicate that this is the Fermi–Dirac distribution function. This function is graphed in Figure 1.9.
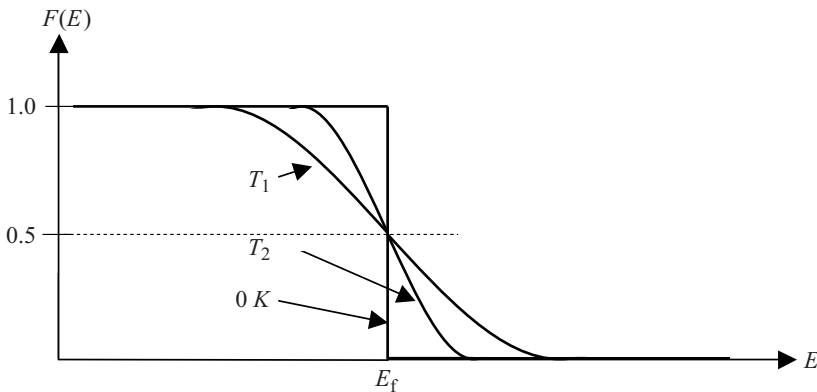
Figure 1.9 *Plot of the Fermi–Dirac distribution function F(E), which gives the probability of occupancy by an electron of an energy state having energy E. The plot is shown for two temperatures $T_1 > T_2$ as well as for 0 K. At absolute zero, the function becomes a step function*
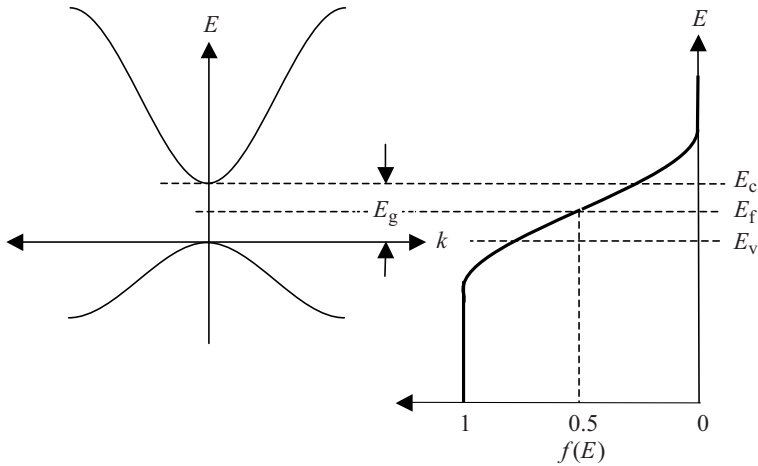
Figure 1.10   *A semiconductor band diagram is plotted along with the Fermi–Dirac distribution function. This shows the probability of occupancy of electron states in the conduction band as well as the valence band. Hole energies increase in the negative direction along the energy axis. The hole having the lowest possible energy occurs at the top of the valence band. This occurs because by convention the energy axis represents electron energies and not hole energies. The origin of the energy axis is located at E_v for convenience*

$F(E)$ is 0.5 at $E = E_f$ provided $T > 0$ K, and at high temperatures the transition becomes more gradual due to increased thermal activation of electrons from lower energy levels to higher energy levels. Figure 1.10 shows $F(E)$ plotted beside a semiconductor band diagram with the energy axis in the vertical direction. The bottom of the conduction band is at E_c and the top of the valence band is at E_v. At E_f there are no electron states since it is in the energy gap; however, above E_c and below E_v the values of $F(E)$ indicate the probability of electron occupancy in the bands. In the valence band the probability for a hole to exist at any energy level is $1 - F(E)$.

The distribution of available energy levels in an energy band is found by knowing the *density of states* function $D(E)$, which gives the number of available energy states per unit volume over a differential energy range. It is needed in order to calculate the number of carriers in an energy band. Knowing the probability of occupancy of the states in a band is not sufficient; the density of available energy levels is also required. Once we have all this information we can obtain the total number of electrons in a band by summing up all the electrons in each energy level within the energy band. The probability of each energy level being occupied by an electron is taken into account in the summation.

The density of states function may be derived by solving Schrödinger's equation for an infinite-walled potential box in which the wavefunctions (Equation 1.17) must be expressed in three dimensions. In three dimensions, Schrödinger's equation is:

$$-\frac{\hbar^2}{2m}\left(\frac{d^2}{dx^2} + \frac{d^2}{dy^2} + \frac{d^2}{dz^2}\right)\psi(x, y, z) + V(x, y, z)\psi(x, y, z) = E\psi(x, y, z)$$

Consider a box of dimensions a, b and c in three-dimensional space in which $V = 0$ inside the box when $0 < x < a, 0 < y < b, 0 < z < c$. Outside the box, assume $V = \infty$.

Inside the box using Schrödinger's equation:

$$-\frac{\hbar^2}{2m}\left(\frac{d^2}{dx^2}+\frac{d^2}{dy^2}+\frac{d^2}{dz^2}\right)\psi(x,y,z)=E\psi(x,y,z) \qquad (1.24)$$

If we let $\psi(x,y,z)=X(x)Y(y)Z(z)$ then upon substitution into Equation 1.24 and after dividing by $\psi(x,y,z)$ we obtain:

$$-\frac{\hbar^2}{2m}\left(\frac{1}{X(x)}\frac{d^2X(x)}{dx^2}+\frac{1}{Y(y)}\frac{d^2Y(y)}{dy^2}+\frac{1}{Z(z)}\frac{d^2Z(z)}{dz^2}\right)=E$$

Since each term contains an independent variable, we can apply *separation of variables* and conclude that each term is equal to a constant that is independent of x, y and z.

Now, we have three equations

$$\frac{1}{X(x)}\frac{d^2X(x)}{dx^2}=-C_1 \qquad (1.25a)$$

$$\frac{1}{Y(y)}\frac{d^2Y(y)}{dy^2}=-C_2 \qquad (1.25b)$$

and

$$\frac{1}{Z(z)}\frac{d^2Z(z)}{dz^2}=-C_3 \qquad (1.25c)$$

where

$$E=\frac{\hbar^2}{2m}(C_1+C_2+C_3) \qquad (1.26)$$

The general solution to Equation 1.25a is

$$X(x)=A_1\exp(ikx)+A_2\exp(-ikx) \qquad (1.27)$$

To satisfy boundary conditions such that $X(x)=0$ at $x=0$ and at $x=a$ we obtain

$$X(x)=A\sin k_x x$$

where

$$k_x=\frac{n_x\pi}{a}$$

with n_x a positive integer quantum number and

$$C_1=\left(\frac{n_x\pi}{a}\right)^2$$

Repeating a similar procedure for Equations 1.25b and 1.25c, and using Equation 1.26 we obtain:

$$\psi(x,y,z)=X(x)Y(y)Z(z)=ABC\sin(k_x x)\sin(k_y y)\sin(k_z z)$$

and

$$E=\frac{\hbar^2\pi^2}{2m}\left(\left(\frac{n_x}{a}\right)^2+\left(\frac{n_y}{b}\right)^2+\left(\frac{n_z}{c}\right)^2\right) \qquad (1.28)$$
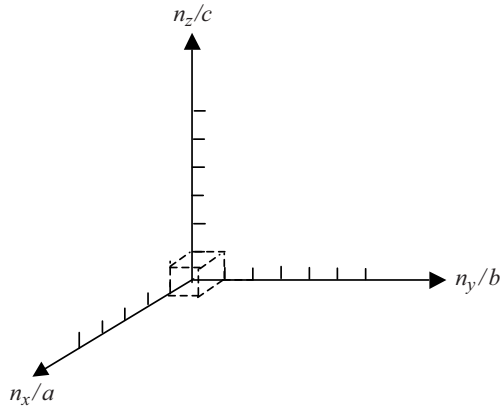
Figure 1.11   *Reciprocal space lattice. A cell in this space is shown, which is the volume associated with one lattice point. The cell has dimensions $\frac{1}{a}$, $\frac{1}{b}$, $\frac{1}{c}$ and volume $\frac{1}{abc}$. This space may readily be transformed into k-space by multiplying each axis by a factor of 2π. It may also readily be transformed into p-space (momentum space) by multiplying each axis by a factor of* h

If more than one electron is put into the box at zero kelvin the available energy states will be filled such that the lowest energy states are filled first.

We now need to determine how many electrons can occupy a specific energy range in the box. It is very helpful to define a three-dimensional space with coordinates $\frac{n_x}{a}$, $\frac{n_y}{b}$ and $\frac{n_z}{c}$. In this three-dimensional space there are discrete points that are defined by these coordinates with integer values of n_x, n_y and n_z in what is referred to as a *reciprocal space lattice*, which is shown in Figure 1.11. Note that reciprocal space is related to another important conceptual space known as *k-space* by multiplying each axis of reciprocal space by a factor of 2π.

From Equation 1.28 it is seen that an ellipsoidal shell in reciprocal space represents an equal energy surface because the general form of this equation is that of an ellipsoid in reciprocal space. The number of reciprocal lattice points that are contained inside the positive octant of an ellipsoid having a volume corresponding to a specific energy E will be the number of states smaller than E. The number of electrons is actually twice the number of these points because electrons have an additional quantum number s for spin and $s = \pm\frac{1}{2}$. The positive octant of the ellipsoid is illustrated in Figure 1.12.

Rearranging Equation 1.28,

$$\left(\frac{n_x}{a}\right)^2 + \left(\frac{n_y}{b}\right)^2 + \left(\frac{n_z}{c}\right)^2 = \frac{2mE}{\hbar^2\pi^2} \tag{1.29}$$

The number of reciprocal lattice points inside the ellipsoid is the volume of the ellipsoid divided by the volume associated with each lattice point shown in Figure 1.11. The volume of the ellipsoid is

$$V = \frac{4}{3}\pi \text{ (product of semi-axes of ellipsoid)}.$$
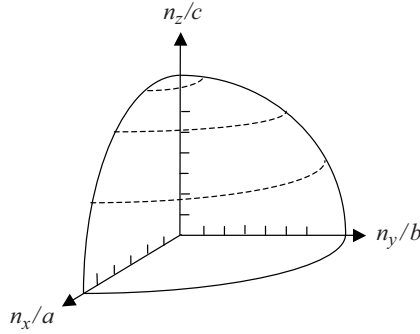
Figure 1.12   *The positive octant of an ellipsoid of revolution in reciprocal space corresponding to an equal energy surface. The number of electron states below this energy is twice the number of reciprocal lattice points inside the positive octant of the ellipsoid*

The semi-axis of length L_x along the $\frac{n_x}{a}$ direction is obtained by setting $n_y = n_z = 0$ and solving for $\frac{n_x}{a}$ in Equation 1.29, and we obtain

$$L_x = \sqrt{\frac{2mE}{\hbar^2\pi^2}}$$

By repeating this for each semi-axis we obtain:

$$V = \frac{4}{3}\pi \left(\frac{2mE}{\hbar^2\pi^2}\right)^{\frac{3}{2}}$$

Now if the volume of the ellipsoid is much larger than the volume associated with one lattice point then, including spin, the number of electrons having energy less than E approaches two times one-eighth of the volume of the ellipsoid (positive octant) divided by the volume associated with one lattice point, or:

$$number\ of\ electrons = \frac{2\left(\frac{1}{8}\right)\frac{4}{3}\pi \left(\frac{2mE}{\hbar^2\pi^2}\right)^{\frac{3}{2}}}{\frac{1}{abc}}$$

We define $n(E)$ to be the number of electrons *per unit volume* of the box and therefore

$$n(E) = 2\left(\frac{1}{8}\right)\frac{4}{3}\pi \left(\frac{2mE}{\hbar^2\pi^2}\right)^{\frac{3}{2}}$$

We also define $D(E)$ to be the *density of states* function where

$$D(E) = \frac{dn(E)}{d(E)}$$

and finally we obtain

$$D(E) = \frac{\pi}{2} \left(\frac{2m}{\hbar^2 \pi^2} \right)^{\frac{3}{2}} E^{\frac{1}{2}} \tag{1.30}$$

This form of the density of states function is valid for a box having $V = 0$ inside the box. In an energy band, however, V is a periodic function and the density of states function must be modified. This is easy to do because rather than the parabolic E versus k dispersion relation (Equation 1.10) for free electrons in which the electron mass is m, we simply use the E versus k dependence for an electron near the bottom or top of an energy band as illustrated in Figure 1.7, which may be approximated as parabolic for small values of k but using the appropriate effective mass. It is important to remember that the density of states function is based on a density of available states in k-space or reciprocal space, and that for a certain range of k-values in k-space the corresponding range of energies along the energy axis is determined by the slope of the E versus k graph. The slope of E versus k in a parabolic band depends on the effective mass (see Section 1.5). Using Equation 1.10 the relevant dispersion curve is now

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

As a result

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

and the second derivative or curvature of the E versus k curve is constant as required.

As a result the density of states function in a conduction band is given by Equation 1.30, provided the effective mass m^* is used in place of m. The point $E = 0$ should refer to the bottom of the band. We now have

$$D(E) = \frac{\pi}{2} \left(\frac{2m^*}{\hbar^2 \pi^2} \right)^{\frac{3}{2}} E^{\frac{1}{2}} \tag{1.31a}$$

Since E_v is defined as zero as in Figure 1.10 for convenience then the conduction band starts at $E_c = E_g$ and $D(E - E_g)$ tells us the number of energy states available per differential range of energy within the conduction band, and we obtain

$$D(E - E_g) = \frac{\pi}{2} \left(\frac{2m_e^*}{\hbar^2 \pi^2} \right)^{\frac{3}{2}} (E - E_g)^{\frac{1}{2}} \tag{1.31b}$$

The total number of electrons per unit volume in the band is now given by

$$n = \int_{E_g}^{E_{max}} D(E - E_g) F(E) dE \tag{1.32}$$

where E_{max} is the highest energy level in the energy band that needs to be considered as higher energy levels have a negligible chance of being occupied.

The integral may be solved analytically provided the upper limit of the integral is allowed to be infinity. This is justifiable because the argument of the integrand is virtually zero above E_{max}.

From Equation 1.23, since $E \geq E_g$ and $E_g - E_f \gg kT$, we can use the *Boltzmann approximation*:

$$F(E) \simeq \exp\left[\frac{-(E - E_f)}{kT}\right] \tag{1.33}$$

Hence from Equations 1.31, 1.32 and 1.33,

$$n = \frac{\pi}{2}\left(\frac{2m_e^*}{\hbar^2\pi^2}\right)^{\frac{3}{2}} \exp\left[-\frac{E_f}{kT}\right] \int_{E_g}^{E_{max}} (E - E_g)^{\frac{1}{2}} \exp\left[-\frac{E}{kT}\right] dE$$

$$= \frac{\pi}{2}\left(\frac{2m_e^*}{\hbar^2\pi^2}\right)^{\frac{3}{2}} \exp\left[-\frac{E_g - E_f}{kT}\right] \int_0^\infty (E)^{\frac{1}{2}} \exp\left[-\frac{E}{kT}\right] dE$$

From standard integral tables and because $E_c = E_g$ we obtain

$$n_0 = N_c \exp\left(\frac{-(E_c - E_f)}{kT}\right) \tag{1.34a}$$

where

$$N_c = 2\left(\frac{2\pi m_e^* kT}{h^2}\right)^{\frac{3}{2}} \tag{1.34b}$$

Here m_e^* is the effective mass of electrons near the bottom of the conduction band. The subscript on n indicates that equilibrium conditions apply. The validity of Equation 1.34 is maintained regardless of the choice of the origin on the energy axis since from Equation 1.34 the important quantity for determining the electron concentration is the *energy difference* between the conduction band edge and the Fermi energy.

The same procedure may be applied to the valence band. In this case we calculate the number of holes p in the valence band. The density of states function must be written as $D(-E)$ since from Figure 1.10 energy E is negative in the valence band and hole energy increases as we move in the negative direction along the energy axis. We can use Equation 1.31a to obtain

$$D(-E) = \frac{\pi}{2}\left(\frac{2m^*}{\hbar^2\pi^2}\right)^{\frac{3}{2}} (-E)^{\frac{1}{2}}$$

The probability of the existence of a hole is $1 - F(E)$, and from Equation 1.23 if $E_f - E \gg kT$ we obtain

$$1 - F(E) \cong \exp\left(\frac{E - E_f}{kT}\right)$$

and now

$$p = \int_0^{-E_{max}} D(-E)(1 - F(E)) dE$$

In an analogous manner to that described for the conduction band, we therefore obtain

$$p_0 = N_v \exp\left(\frac{-(E_f - E_v)}{kT}\right) \tag{1.35a}$$

where

$$N_v = 2 \left(\frac{2\pi m_h^* kT}{h^2} \right)^{\frac{3}{2}} \tag{1.35b}$$

and m_h^*, the hole effective mass, is a positive quantity.

Equation 1.35 shows that the important quantity for the calculation of hole concentration in Equation 1.16 is the *energy difference* between the Fermi energy and the valence band edge. Again the subscript on p indicates that equilibrium conditions apply.

We can now determine the position of the Fermi level and will again set $E_v = 0$ for convenience as illustrated in Figure 1.10. Since $n_i = p_i$ for an intrinsic semiconductor we equate Equations 1.34 and 1.35 and obtain

$$N_c \exp \left(\frac{-(E_g - E_f)}{kT} \right) = N_v \exp \left(\frac{-E_f}{kT} \right)$$

or

$$E_f = \frac{E_g}{2} + \frac{kT}{2} \ln \frac{N_v}{N_c} \tag{1.36}$$

The second term on the right side of Equation 1.36 is generally much smaller than $\frac{E_g}{2}$ and we conclude that the Fermi energy lies approximately in the middle of the energy gap.

From Equations 1.34 and 1.35 we can also write the product

$$np = N_c N_v \exp \left(\frac{-E_g}{kT} \right) \tag{1.37a}$$

and for an intrinsic semiconductor with $n_i = p_i$

$$n_i = p_i = \sqrt{N_c N_v} \exp \left(\frac{-E_g}{2kT} \right) \tag{1.37b}$$

which is a useful expression for carrier concentration as it is independent of E_f.

Example 1.2

(a) Calculate $n_i = p_i$ for silicon at room temperature and compare with the commonly accepted value.
(b) Calculate $n_i = p_i$ for gallium arsenide at room temperature.

Solution

(a) Using Appendix 2 to obtain silicon values $m_e^* = 1.08m$ and $E_g = 1.11$ eV,

$$N_c = 2 \left(\frac{2\pi m_e^* kT}{h^2} \right)^{\frac{3}{2}} = 2 \left(\frac{2\pi \times (1.08 \times 9.11 \times 10^{-31} \text{ kg}) \times (0.026 \times 1.6 \times 10^{-19} \text{ J})}{(6.625 \times 10^{-34} \text{ J s})^2} \right)^{\frac{3}{2}}$$

$$= 2.84 \times 10^{25} \text{ m}^{-3} = 2.84 \times 10^{19} \text{ cm}^{-3}$$

and

$$N_{\mathrm{v}} = 2\left(\frac{2\pi m_{\mathrm{h}}^* kT}{h^2}\right)^{\frac{3}{2}} = 2\left(\frac{2\pi \times (0.56 \times 9.11 \times 10^{-31}\ \mathrm{kg}) \times (0.026 \times 1.6 \times 10^{-19}\mathrm{J})}{(6.625 \times 10^{-34}\mathrm{J\ s})^2}\right)^{\frac{3}{2}}$$

$$= 1.06 \times 10^{25}\ \mathrm{m}^{-3} = 1.06 \times 10^{19}\ \mathrm{cm}^{-3}$$

Now,

$$n_{\mathrm{i}} = p_{\mathrm{i}} = \sqrt{N_{\mathrm{c}} N_{\mathrm{v}}}\ \exp\left(\frac{-E_{\mathrm{g}}}{2kT}\right) = \sqrt{2.84 \times 10^{19} \times 1.06 \times 10^{19}}$$

$$\times\ \exp\left(\frac{-1.11\ \mathrm{eV}}{2 \times 0.026\ \mathrm{eV}}\right) = 9.31 \times 10^9 \mathrm{cm}^{-3}$$

The commonly accepted value is $n_{\mathrm{i}} = p_{\mathrm{i}} = 1.5 \times 10^{10}\ \mathrm{cm}^{-3}$. The discrepancy relates mainly to three-dimensional aspects of the effective mass value, and the method and temperature at which effective mass is measured. We will continue to use the commonly accepted effective mass unless otherwise noted.

(b) For GaAs from Appendix 2, $m_{\mathrm{e}}^* = 0.067$ and $E_{\mathrm{g}} = 1.42$ eV. Hence

$$N_{\mathrm{c}} = 2\left(\frac{2\pi m_{\mathrm{e}}^* kT}{h^2}\right)^{\frac{3}{2}} = 2\left(\frac{2\pi \times (0.067 \times 9.11 \times 10^{-31}\ \mathrm{kg}) \times (0.026 \times 1.6 \times 10^{-19}\ \mathrm{J})}{(6.625 \times 10^{-34}\ \mathrm{J\ s})^2}\right)^{\frac{3}{2}}$$

$$= 4.38 \times 10^{23}\ \mathrm{m}^{-3} = 4.38 \times 10^{17}\ \mathrm{cm}^{-3}$$

and

$$N_{\mathrm{v}} = 2\left(\frac{2\pi m_{\mathrm{h}}^* kT}{h^2}\right)^{\frac{3}{2}} = 2\left(\frac{2\pi \times (0.48 \times 9.11 \times 10^{-31}\ \mathrm{kg}) \times (0.026 \times 1.6 \times 10^{-19}\ \mathrm{J})}{(6.625 \times 10^{-34}\ \mathrm{J\ s})^2}\right)^{\frac{3}{2}}$$

$$= 8.4 \times 10^{24}\ \mathrm{m}^{-3} = 8.4 \times 10^{18}\ \mathrm{cm}^{-3}$$

Now,

$$n_{\mathrm{i}} = p_{\mathrm{i}} = \sqrt{N_{\mathrm{c}} N_{\mathrm{v}}}\ \exp\left(\frac{-E_{\mathrm{g}}}{2kT}\right) = \sqrt{4.38 \times 10^{17} \times 8.4 \times 10^{18}}\ \exp\left(\frac{-1.42\ \mathrm{eV}}{2 \times 0.026\ \mathrm{eV}}\right)$$

$$= 2.65 \times 10^6\ \mathrm{cm}^{-3}$$

1.10   Semiconductor Materials

The relationship between carrier concentration and E_{g} has now been established and we can look at examples of real semiconductors. A portion of the periodic table showing elements from which many important semiconductors are made is shown in Figure 1.13, together with a list of selected semiconductors and their energy gaps. Note that there are the group IV semiconductors silicon and germanium, a number of III-V compound semiconductors

Group	II	III	IV	V	VI
		B	C	N	O
Element		Al	Si	P	S
	Zn	Ga	Ge	As	Se
	Cd	In	Sn	Sb	Te

Group	IV	IV	IV	IV	III-V	III-V	III-V	III-V	III-V	III-V	II-VI	II-VI
Element(s)	C	Si	Ge	Sn	GaN	AlP	GaP	AlAs	GaAs	InSb	ZnSe	CdTe
Energy gap (eV)	6	1.11	0.67	0	3.4	2.45	2.26	2.16	1.43	0.18	2.7	1.58

Figure 1.13   *A portion of the periodic table containing some selected semiconductors composed of elements in groups II to VI*

having two elements, one from group III and one from group V respectively, and a number of II-VI compound semiconductors having elements from group II and group VI respectively.

A number of interesting observations may now be made. In group IV crystals, the bonding is purely covalent. Carbon (diamond) is an insulator because it has an energy gap of 6 eV. The *energy gap decreases with atomic size* as we look down the group IV column from C to Si to Ge and to Sn. Actually Sn behaves like a metal. Since its energy gap is very small, it turns out that the valence band and conduction band effectively overlap when a three-dimensional model of the crystal is considered rather than the one-dimensional model we have discussed. This guarantees some filled states in the conduction band and empty states in the valence band regardless of temperature. Sn is properly referred to as a *semi-metal* (its conductivity is considerably lower than metals like copper or silver). We can understand this group IV trend of decreasing energy gaps since the periodic potential of heavy elements will be weaker than that of lighter elements due to electron screening as described in Section 1.3.

As with group IV materials, the energy gaps of III-V semiconductors decrease as we go down the periodic table from AlP to GaP to AlAs to GaAs and to InSb. The energy gaps of II-VI semiconductors behave in the same manner as illustrated by ZnSe and CdTe. Again, electron screening increases for heavier elements.

If we compare the energy gaps of a set of semiconductors composed of elements from the same row of the periodic table but with increasingly ionic bonding such as Ge, GaAs and ZnSe, another trend becomes clear: *Energy gaps increase as the degree of ionic character becomes stronger.* The degree of ionic bond character increases the magnitude of the periodic potential and hence the energy gap.

The carrier concentration as a function of temperature according to Equation 1.18 is plotted for three semiconductors in Figure 1.14. Increasing energy gaps result in lower carrier concentrations at a given temperature.

1.11   Semiconductor Band Diagrams

The semiconductors in Figure 1.10 crystallize in either cubic or hexagonal structures. Figure 1.15a shows the *diamond* structure of silicon, germanium (and carbon), which is cubic.
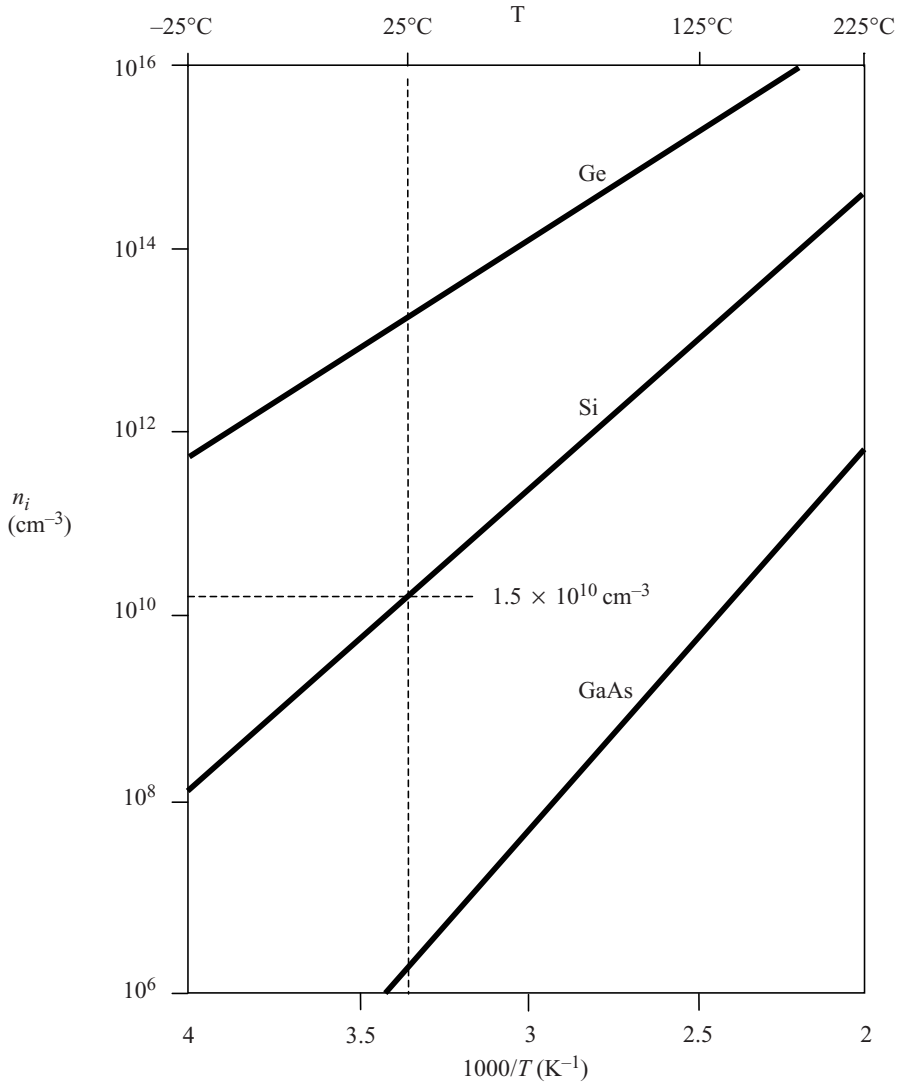
Figure 1.14 *Plot of commonly accepted values of* n *as a function of* $\frac{1}{T}$ *for intrinsic germanium* $(E_g = 0.7\ eV)$, *silicon* $(E_g = 1.1\ eV)$, *and gallium arsenide* $(E_g = 1.43\ eV)$

Figure 1.15b shows the *zincblende* structure of a set of III-V and II-VI semiconductors, which is also cubic. Figure 1.15c shows the *hexagonal* structure of some additional compound semiconductors.

   These three structures have features in common. Each atom has four nearest neighbours in a tetrahedral arrangement. Some crystals exhibit distortions from the ideal 109.47-degree tetrahedral bond angle; however, since all the compounds have directional covalent bonding
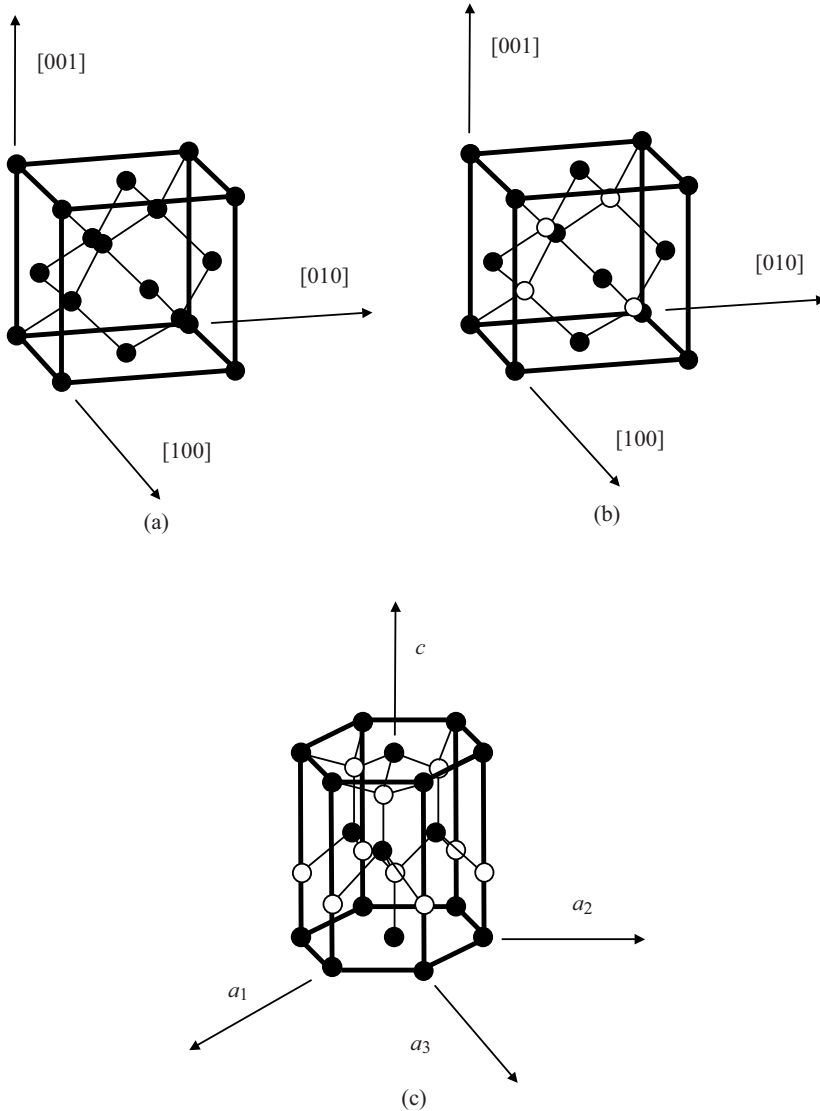
Figure 1.15   *(a) The diamond unit cell of crystal structures of C, Si and Ge. The cubic unit cell contains eight atoms. Each atom has four nearest neighbours in a tetrahedral arrangement. Within each unit cell, four atoms are arranged at the cube corners and at the face centres in a face-centred cubic (FCC) sublattice, and the other four atoms are arranged in another FCC sublattice that is offset by a translation along one quarter of the body diagonal of the unit cell. (b) The zincblende unit cell contains four 'A' atoms (black) and four 'B' atoms (white). The 'A' atoms form an FCC sublattice and the 'B' atoms form another FCC sublattice that is offset by a translation along one quarter of the body diagonal of the unit cell. (c) The hexagonal wurtzite unit cell contains six 'A' atoms and six 'B' atoms. The 'A' atoms form a hexagonal close-packed (HCP) sublattice and the 'B' atoms form another HCP sublattice that is offset by a translation along the vertical axis of the hexagonal unit cell. Each atom is tetrahedrally bonded to four nearest neighbours. A vertical axis in the unit cell is called the c-axis*

to some degree, bond angles do not vary widely. Both the cubic (111) planes and the wurtzite (1000) planes normal to the *c*-axis have close-packed hexagonal atomic arrangements.

The energy gap and effective mass values for a given semiconductor are not sufficient information for optoelectronic applications. We need to re-examine the energy band diagrams for real materials in more detail.

The Kronig–Penney model involves several approximations. A one-dimensional periodic potential instead of a three-dimensional periodic potential is used. The periodic potential is simplified, and does not actually replicate the atomic potentials in real semiconductor crystals. For example, silicon has a diamond crystal structure with silicon atoms as shown in Figure 1.15a. Not only are three dimensions required, but also there is more than one atom per unit cell.

In addition charges associated with individual atoms in compound semiconductors depend on the degree of ionic character in the bonding. This will affect the detailed shape of the periodic potential. Also effects of electron shielding have not been accurately modelled. There are also other influences from electron spin and orbital angular momentum that influence energy bands in real crystals.

E versus *k* diagrams for various directions in a semiconductor crystal are often presented since the one-dimensional periodic potentials vary with direction. Although three-dimensional modelling is beyond the scope of this book the results for cubic crystals of silicon, germanium, gallium arsenide, gallium phosphide, gallium nitride and cadmium telluride as well as for wurtzite GaN are shown in Figure 1.16a–g. For cubic crystals these figures show the band shape for an electron travelling in the [111] crystal direction on the left side and for the [100] direction on the right side. It is clear that the periodic potential experienced by an electron travelling in various directions changes: the value of *a* in $u_k(x) = u_k(x + a)$ appropriate for use in the Bloch function (Equation 1.5) for the [100] direction is the edge length of the cubic unit cell of the crystal. For the [111] direction *a* must be modified to be the distance between the relevant atomic planes normal to the body diagonal of the unit cell. For wurtzite crystals the two directions shown are the [0001] direction along the *c*-axis and the ⟨1100⟩ directions along the *a*-axes.

Note that there are multiple valence bands that overlap or almost overlap with each other rather than a single valence band. These are *sub-bands* for holes, which are due to spin–orbit interactions that modify the band state energies for electrons in the valence band. The sub-bands are approximately parabolic near their maxima. Because the curvatures of these sub-bands vary, they give rise to what are referred to as *heavy holes* and *light holes* with m^* as described by Equation 1.16. There are also *split-off* bands with energy maxima below the valence band edge.

1.12   Direct Gap and Indirect Gap Semiconductors

In Figure 1.16 the conduction bands generally exhibit two energy minima rather than one minimum. Each local minimum can be approximated by a parabola whose curvature will determine the effective mass of the relevant electrons.

Referring to Figure 1.16c, we can see that the bandgap of GaAs is 1.43 eV where the valence band maximum and conduction band minimum coincide at $k = 0$. This occurs because the overall minimum of the conduction band is positioned at the same value of

The band structure diagram with:

Energy   300 K $E_g = 1.1$ eV
                                       $E_L = 2.0$ eV
                                       $E_X = 1.2$ eV
                                       $E_{so} = 0.044$ eV
                                       $E_{\Gamma 1} = 3.4$ eV
                                       $E_{\Gamma 2} = 4.2$ eV

$E_{\Gamma 2}$ $E_{\Gamma 1}$ E_L

<100>   <111>
                                       Wave vector

E_X E_g

E_{so} Heavy holes

                            Light holes

Split-off band
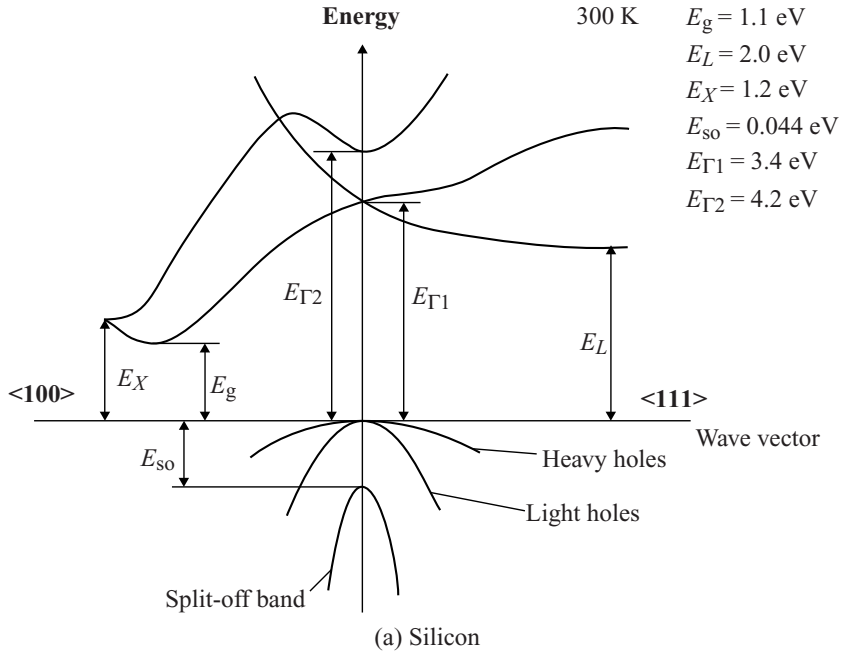
(a) Silicon

Figure 1.16 *Band structures of selected semiconductors. (a) silicon, (b) germanium, (c) GaAs, (d) GaP, (e) cubic GaN, (f) CdTe and (g) wurtzite GaN. Note that GaN is normally wurtzite. Cubic GaN is not an equilibrium phase at atmospheric pressure;however, it can be prepared at high pressure and it is stable once grown. Note that symbols are used to describe various band features. Γ denotes the point where* k = 0. X *and* L *denote the Brillouin zone boundaries in the ⟨100⟩ and ⟨111⟩ directions respectively in a cubic semiconductor. In (g)* k_x *and* k_z *denote the* a *and* c *directions respectively in a hexagonal semiconductor. See Figure 1.15c. Using the horizontal axes to depict two crystal directions saves drawing an additional figure; it is unnecessary to show the complete drawing for each* k*-direction since the positive and negative* k*-axes for a given* k*-direction are symmetrical. There are also energy gaps shown that are larger than the actual energy gap; the actual energy gap is the smallest gap. These band diagrams are the result of both measurements and modelling results. In some cases the energy gap values differ slightly from the values in Appendix 2. (a–d) Reprinted from Levinstein, M., Rumyantsev, S., and Shur, M., Handbook Series on Semiconductor Parameters vol. 1. ISBN 9810229348. Copyright (1996) with permission from World Scientific, London. (e, g) Reprinted from Morkoc, H., Handbook of Nitride Semiconductors and Devices, Vol. 1, ISBN 978-3-527-40837-5. Copyright (2008) WILEY-VCH Verlag GmbH & Co. KGaA Weinheim. (f) Reprinted from Chadov, S., et al., Tunable multifunctional topological insulators in ternary Heusler compounds, Nature Materials 9, 541–545. Copyright (2010) with permission from Nature Publishing Group. DOI: doi:10.1038/nmat2770*

(b) Germanium

300 K E_g = 0.66 eV
E_X = 1.2 eV
$E_{\Gamma 1}$ = 0.8 eV
$E_{\Gamma 2}$ = 3.22 eV
ΔE = 0.85 eV
E_{so} = 0.29 eV

(c) Gallium Arsenide

300 K E_g = 1.43 eV
E_L = 1.71 eV
E_X = 1.90 eV
E_{so} = 0.34 eV

Figure 1.16   *(Continued)*

(d) Gallium Phosphide

(e) Gallium Nitride (cubic)
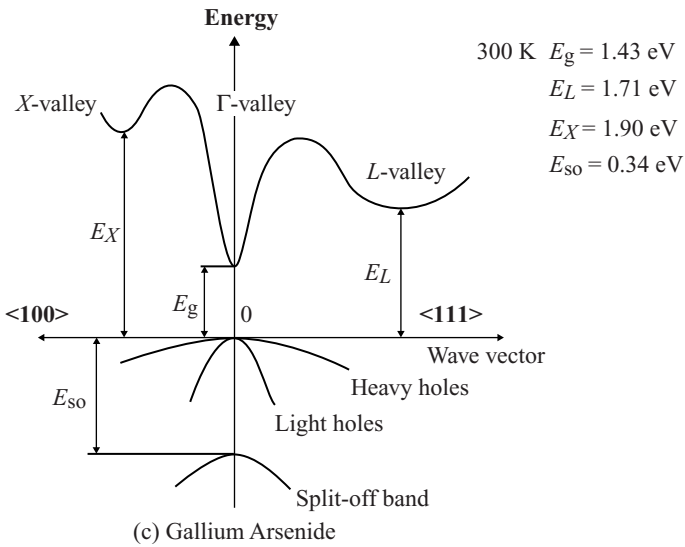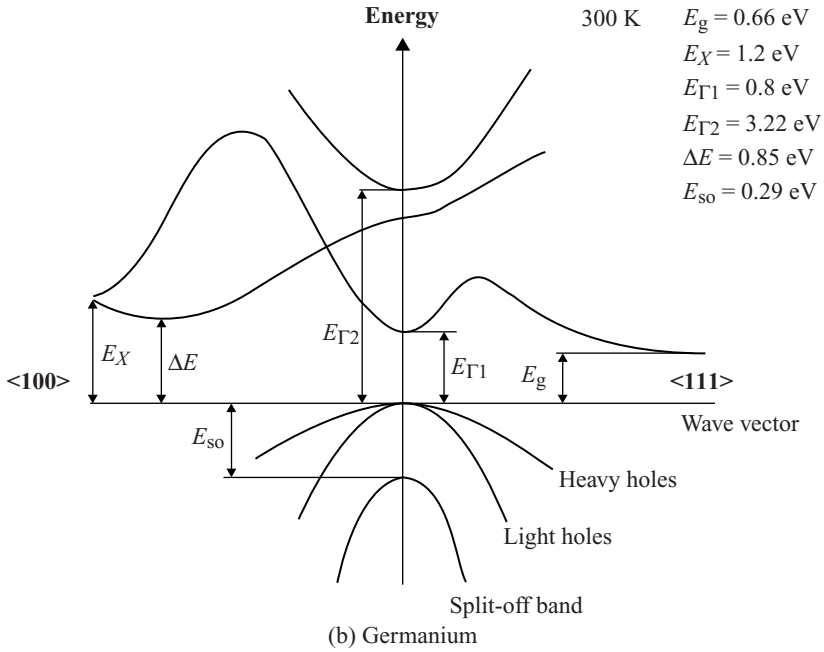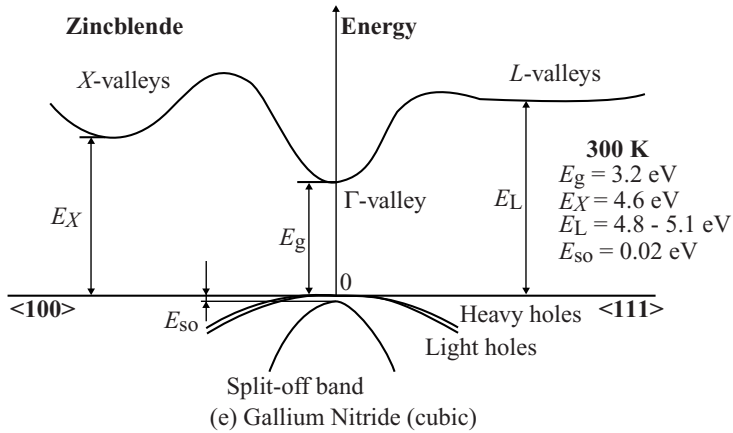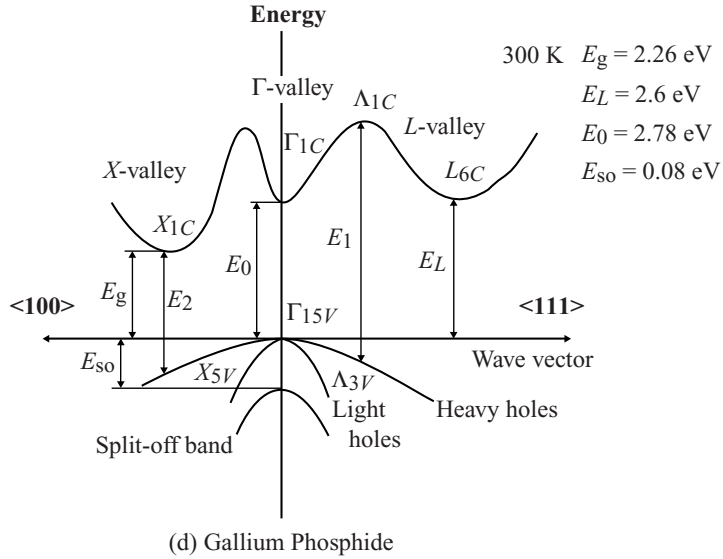
Figure 1.16   *(Continued)*

k as the valence band maximum and this results in a *direct gap* semiconductor. In Figure 1.16 GaAs, GaN and CdTe are direct gap semiconductors. In contrast to GaAs, silicon in Figure 1.16a has a valence band maximum at a different value of k than the conduction band minimum. That means that the energy gap of 1.1 eV is not determined by the separation between bands at $k = 0$, but rather by the distance between the overall conduction band minimum and valence band maximum. This results in an *indirect gap* semiconductor. Another indirect gap semiconductor in Figure 1.16 is the III-V material GaP.

(f) Cadmium Telluride

$E_g = 1.58$ eV

Energy (eV)

CdTe

L Γ X

Wurtzite

Energy

A-valley

M-L-valleys

Γ-valley

300 K
$E_g = 3.39$ eV
$E_{M-L} = 4.5 - 5.3$ eV
$E_A = 4.7 - 5.5$ eV
$E_{so} = 0.008$ eV
$E_{cr} = 0.04$ eV

E_A

E_g

E_{M-L}

k_z E_{cr} Heavy holes k_x
                                        Light holes

Split-off band
(g) Gallium Nitride (wurtzite)

Figure 1.16   *(Continued)*

The distinction between direct and indirect gap semiconductors is of particular significance for photovoltaic and LED devices because processes involving photons occur in both cases, and photon absorption and generation properties differ considerably between these two semiconductor types.

An *electron-hole pair* (EHP) may be created if a photon is absorbed by a semiconductor and causes an electron in the valence band to be excited into the conduction band. For example, photon absorption in silicon can occur if the photon energy matches or exceeds the bandgap energy of 1.11 eV. Since silicon is an indirect gap semiconductor, however, there is a shift along the k axis for the electron that leaves the top of the valence band and then occupies the bottom of the conduction band. In Section 1.3 we noted that $p = \hbar k$ and therefore a shift in momentum results. The shift is considerable as seen in Figure 1.16a, and it is almost the distance from the centre of the Brillouin zone at $k = 0$ to the zone boundary at $k = \frac{\pi}{a}$ yielding a momentum shift of

$$\Delta p \simeq \hbar \frac{\pi}{a} \qquad (1.38)$$

During the creation of an EHP both energy and momentum must be conserved. Energy is conserved since the photon energy $\hbar\omega$ satisfies the condition $\hbar\omega = E_g$. Photon momentum $p = \frac{h}{\lambda}$ is very small, however, and is unable to provide momentum conservation. This is discussed further in Section 4.2. This means that a lattice vibration, or *phonon*, is required to take part in the EHP generation process. The magnitudes of phonon momenta cover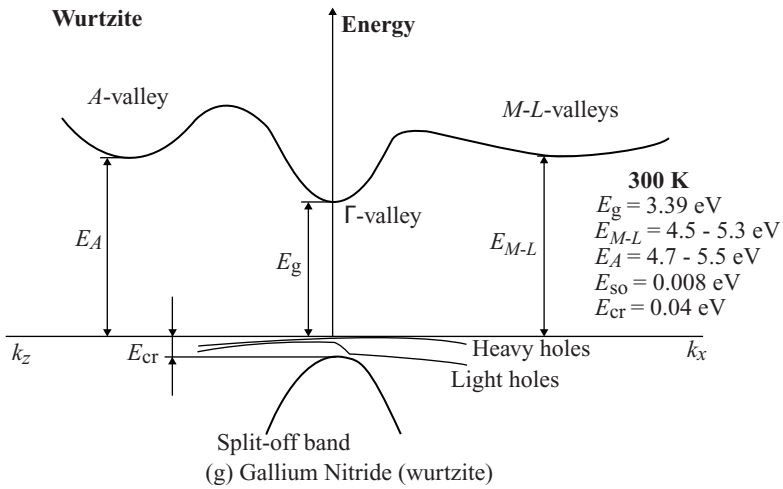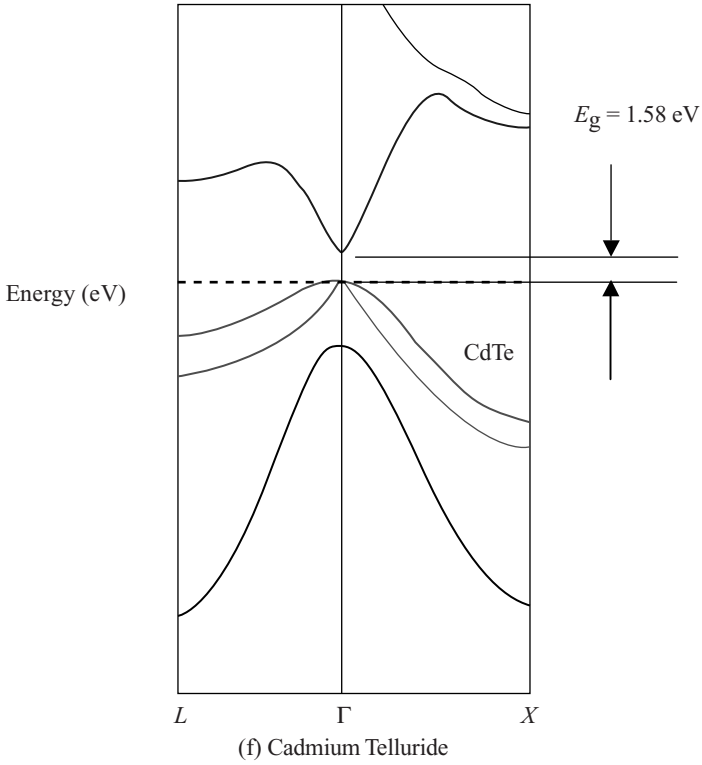 a wide range in crystals and a phonon with the required momentum may not be available to the EHP process. This limits the rate of EHP generation, and photons that are not absorbed continue to propagate through the silicon.

If electromagnetic radiation propagates through a semiconductor we quantify absorption using an *absorption coefficient* α, which determines the intensity of radiation by the exponential relationship

$$\frac{I(x)}{I_0} = e^{-\alpha x}$$

where I_0 is the initial radiation intensity and $I(x)$ is the intensity after propagating through the semiconductor over a distance x. Efficient crystalline silicon solar cells are generally at least $\simeq 100\ \mu m$ thick for this reason due to their relatively low absorption coefficient. In contrast, GaAs (Figure 1.16c), is a direct gap semiconductor and has a much higher value of α (see Section 4.2). The thickness of GaAs required for sunlight absorption is only $\simeq 1\ \mu m$. The value of α is an important parameter in PV semiconductors since sunlight that is not absorbed will not contribute to electric power generation. It is interesting to note that in spite of this difficulty silicon has historically been the most important solar cell material owing to its large cost advantage over GaAs.

In LEDs the process is reversed. EHPs recombine and give rise to photons, which are emitted as radiation. The wavelength range of this radiation may be in the infrared, the visible, or the ultraviolet parts of the electromagnetic spectrum, and is dependent on the semiconductor energy gap. Silicon is a poor material for LEDs because for an EHP recombination to create a photon, one or more phonons need to be involved to achieve momentum conservation. The probability for this to occur is therefore much smaller and competing mechanisms for electron-hole pair recombination become important. These are

known as *non-radiative recombination* events (see Section 1.19). In contrast to silicon, GaAs can be used for high-efficiency LEDs and was used for the first practical LED devices due to its direct gap.

1.13   Extrinsic Semiconductors

The incorporation of very small concentrations of impurities, referred to as *doping*, allows us to create semiconductors that are called *extrinsic* to distinguish them from intrinsic semiconductors, and we can control both the electron and hole concentrations over many orders of magnitude.

Consider the addition of a group V atom such as phosphorus to a silicon crystal as shown together with a band diagram in Figure 1.17. This results in an *n-type semiconductor*. The phosphorus atom substitutes for a silicon atom and is called a *donor*; it introduces a new spatially localized energy level called the donor level E_d.

Because phosphorus has one more electron than silicon this donor electron is not required for valence bonding, is only loosely bound to the phosphorus, and can easily be excited into the conduction band. The energy required for this is $E_c - E_d$ and is referred to as the *donor binding energy*. If the donor electron has entered the conduction band, it is no longer spatially localized and the donor becomes a positively charged ion. The donor binding energy may be calculated by considering the well-known hydrogen energy quantum states in which the ionization energy for a hydrogen atom is given by

$$E_{\text{Rydberg}} = \frac{-mq^4}{8\epsilon_0^2 h^2} = 13.6 \text{ eV} \tag{1.39}$$

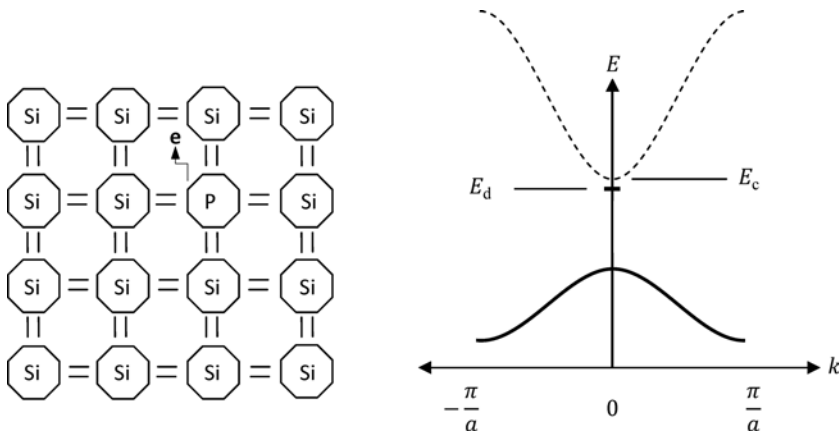**Figure 1.17**   *The substitution of a phosphorus atom in silicon (donor atom) results in a weakly bound extra electron occupying new energy level E_d that is not required to complete the covalent bonds in the crystal. It requires only a small energy $E_c - E_d$ to be excited into the conduction band, resulting in a positively charged donor ion and an extra electron in the conduction band*

and the Bohr radius given by

$$a_0 = \frac{4\pi\epsilon_0 \, \hbar^2}{mq^2} = 0.529 \text{ Å} \qquad (1.40)$$

Now, two variables in Equations 1.39 and 1.40 must be changed. Whereas the hydrogen electron moves in a vacuum, the donor is surrounded by semiconductor atoms, which requires us to modify the dielectric constant from the free space value ϵ_0 to the appropriate value for silicon by multiplying by the relative dielectric constant ϵ_r. In addition the free electron mass m must be changed to the effective mass m_e^*. This results in a small binding energy from Equation 1.39 compared to the hydrogen atom, and a large atomic radius from Equation 1.40 compared to the Bohr radius. For n-type dopants in silicon the measured values of binding energy are approximately 0.05 eV compared to 13.6 eV for the Rydberg constant, and an atomic radius is obtained that is an order of magnitude larger than the Bohr radius of approximately 0.5 Å. Since the atomic radius is now several lattice constants in diameter, we can justify the use of the bulk silicon constants we have used in place of vacuum constants.

Consider now the substitution of a group III atom such as aluminium for a silicon atom as illustrated in Figure 1.18. This creates a *p-type semiconductor*. The aluminium atom is called an *acceptor* and it introduces a new spatially localized energy level called the acceptor level E_a. Because aluminium has one fewer electron than silicon it can accept an electron from another valence bond elsewhere in the silicon, which results in a hole in the valence band. The energy required for this is $E_a - E_v$ and is referred to as the *acceptor binding energy*. If an electron has been accepted, the resulting hole is no longer spatially
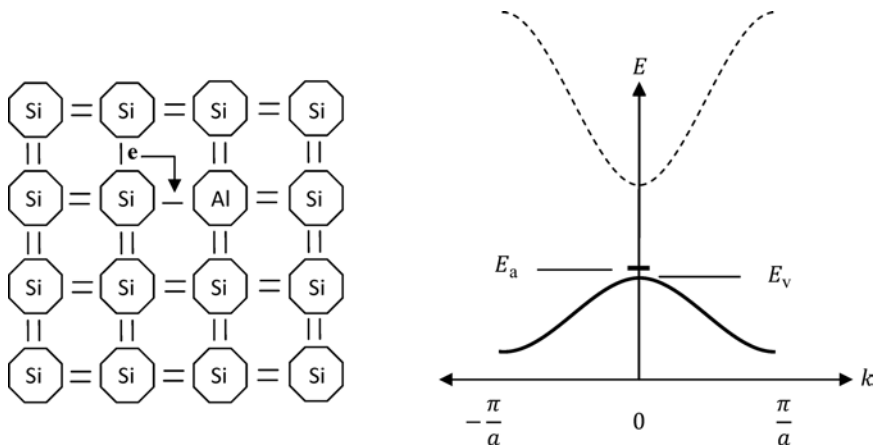
Figure 1.18   *The substitution of an aluminium atom in silicon (acceptor atom) results in an incomplete valence bond for the aluminium atom. An extra electron may be transferred to fill this bond from another valence bond in the crystal. The spatially localized energy level now occupied by this extra electron at E_a is slightly higher in energy than the valence band. This transfer requires only a small energy $E_a - E_v$ and results in a negatively charged acceptor ion and an extra hole in the valence band*

localized and the acceptor becomes a negatively charged ion. The binding energy may be estimated in a manner analogous to donor binding energies.

The introduction of either donors or acceptors influences the concentrations of charge carriers, and we need to be able to calculate these concentrations. The position of the Fermi level changes when dopant atoms are added, and it is no longer true that $n = p$; however, the Fermi–Dirac function $F(E)$ still applies. A very useful expression becomes the product of electron and hole concentrations in a given semiconductor. For intrinsic material, we have calculated $n_i p_i$ and we obtained Equation 1.37a; however, Equations 1.34 and 1.35 are still valid and we can also conclude that

$$n_0 p_0 = n_i^2 = p_i^2 = N_c N_v \exp\left(\frac{-E_g}{kT}\right) \tag{1.41}$$

which is independent of E_f, and therefore is also applicable to extrinsic semiconductors. Here n_0 and p_0 refer to the equilibrium carrier concentrations in the doped semiconductor.

We now examine the *intermediate temperature condition* where the following apply:

(a) The ambient temperature is high enough to ionize virtually all the donors or acceptors.
(b) The concentration of the dopant is much higher than the intrinsic carrier concentration because the ambient temperature is not high enough to directly excite a large number of electron-hole pairs.

Under these circumstances, there are two cases. For donor doping in an n-type semiconductor we can conclude that

$$n_0 \simeq N_d \tag{1.42}$$

and combining Equations 1.41 and 1.42 we obtain

$$p_0 = \frac{n_i^2}{N_d} \tag{1.43}$$

where N_d is the donor concentration in donor atoms per unit volume of the semiconductor. For acceptor doping in a p-type semiconductor we have

$$p_0 \simeq N_a \tag{1.44}$$

and we obtain

$$n_0 = \frac{n_i^2}{N_a} \tag{1.45}$$

The Fermi energy levels will change upon doping, and may be calculated from Equations 1.34 and 1.35. In the case of n-type silicon the Fermi level will lie closer to the conduction band. In the case of p-type silicon the Fermi level will lie closer to the valence band (see Figure 1.19). In Example 1.3 we calculate some specific values of the Fermi energy position.

Consider n-type silicon at room temperature. The mobile electrons in the n-type silicon are called *majority carriers*, and the mobile holes are called *minority carriers*. We can also consider p-type silicon with mobile holes called majority carriers and mobile electrons called minority carriers.
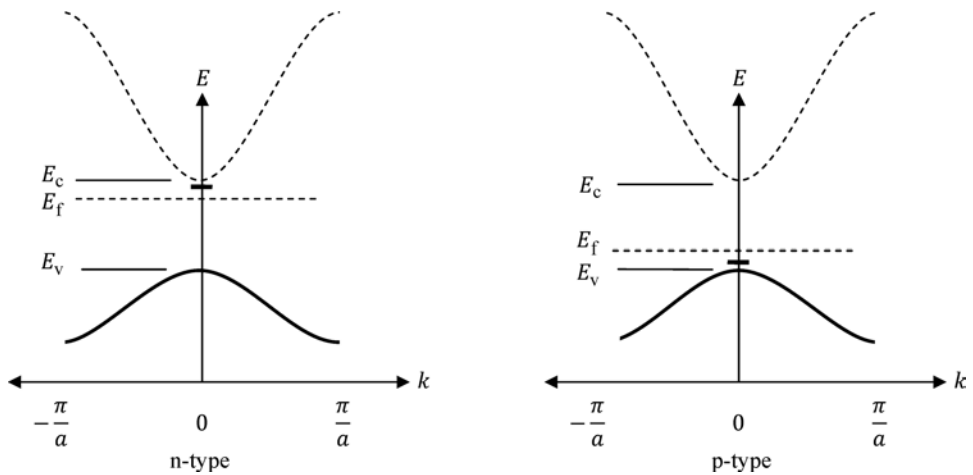
Figure 1.19   *The band diagrams for n-type silicon with a donor doping concentration of 1×10^{17} cm^{-3} and p-type silicon with an acceptor doping concentration of 1×10^{17} cm^{-3}. Note that the Fermi energy rises to the upper part of the energy gap for n-type doping and drops to the lower part of the energy gap for p-type doping*

Example 1.3

Assume a silicon sample at room temperature.

(a) Calculate the separation between E_c and E_f for n-type silicon having a phosphorus impurity concentration of 1×10^{17} cm^{-3}. Find both electron and hole concentrations.
(b) Calculate the separation between E_c and E_f for p-type silicon having an aluminium impurity concentration of 1×10^{17} cm^{-3}. Find both electron and hole concentrations.

Solution

(a) Using the intermediate temperature approximation $n_0 \cong 1 \times 10^{17}$ cm^{-3}. From example 1.2 $n_i = 1.5 \times 10^{10}$ cm^3 and hence

$$p_0 = \frac{n_i^2}{n_0} = \frac{(1.5 \times 10^{10} \text{ cm}^{-3})^2}{1 \times 10^{17} \text{ cm}^{-3}} = 2.25 \times 10^3 \text{ cm}^{-3}$$

and

$$n_0 = N_c \exp\left(\frac{-(E_c - E_f)}{kT}\right)$$

Solving for $E_c - E_f$,

$$E_c - E_f = kT \ln \frac{N_c}{n_0} = 0.026 \text{ eV} \; \ln \left(\frac{2.84 \times 10^{19} \text{ cm}^{-3}}{1 \times 10^{17} \text{ cm}^{-3}} \right) = 0.15 \text{ eV}$$
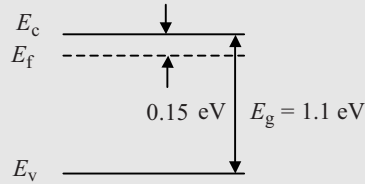
(b) Using the intermediate temperature approximation, $p_0 \cong 1 \times 10^{17} \text{ cm}^{-3}$ and hence

$$n_0 = \frac{n_i^2}{p_0} = \frac{(1.5 \times 10^{10} \text{ cm}^{-3})^2}{1 \times 10^{17} \text{ cm}^{-3}} = 2.25 \times 10^3 \text{ cm}^{-3}$$

and

$$p_0 = N_v \exp \left(\frac{-(E_f - E_v)}{kT} \right)$$

Solving for $E_f - E_v$

$$E_f - E_v = kT \ln \frac{N_v}{n_0} = 0.026 \text{ eV} \; \ln \left(\frac{1.06 \times 10^{19} \text{ cm}^{-3}}{1 \times 10^{17} \text{ cm}^{-3}} \right) = 0.12 \text{ eV}$$
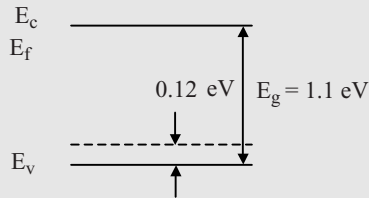
At low ambient temperatures, the dopant atoms are not necessarily ionized. In this case there will be a combination of neutral and ionized dopant atoms. At high ambient temperatures, the intrinsic electron-hole pair concentration may be significant and may exceed the doping concentration. In this case the semiconductor carrier concentrations can be similar to intrinsic material. These cases are illustrated in Figure 1.20. Of particular technological importance is the intermediate temperature region since the carrier concentrations are relatively independent of temperature and therefore semiconductor devices can operate over wide temperature ranges without significant variation in carrier concentrations.
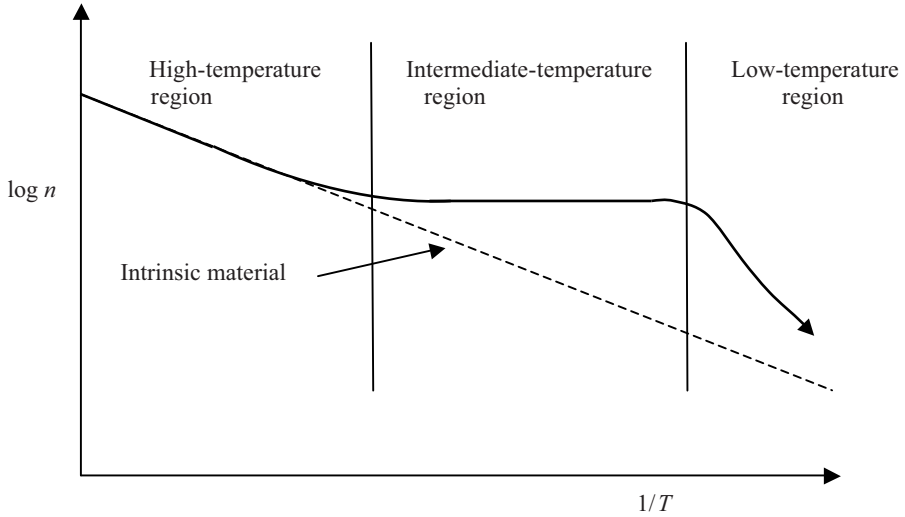
Figure 1.20   *Carrier concentration as a function of temperature for an n-type extrinsic semi-conductor. In the high-temperature region behaviour is intrinsic-like. In the intermediate-temperature region carrier concentration is controlled by the impurity concentration and virtually all the dopant atoms are ionized. At low temperatures there is not enough thermal energy to completely ionize the dopant atoms*

1.14   Carrier Transport in Semiconductors

The electrical conductivity of semiconductors is controlled by the concentrations of both holes and electrons as well as their ability to flow in a specific direction under the influence of an electric field. The flow of carriers is limited by scattering events in which carriers having a high instantaneous velocity frequently scatter off lattice vibrations (phonons), defects and impurities, and we can denote a *scattering time* or characteristic mean time between scattering events for this, referred to as τ. The resulting net flow velocity or *drift velocity* of a stream of carriers is much lower than their instantaneous velocity. The experimental evidence for this is summarized by Ohm's law, or

$$J = \sigma \varepsilon$$

which is a collision-limited flow equation that relates the current flow to the applied electric field. To understand this we consider Figure 1.21 showing the flow of carriers in a solid cylinder of cross-sectional area A in the x direction.

If the carrier concentration is n and each carrier carrying charge q moves a distance dx in time dt then the amount of charge dQ passing across a given plane in the cylinder in time dt is $dQ = nqAdx$. The carrier drift velocity is given by $\bar{v} = \frac{dx}{dt}$, and we can conclude that the current is

$$I = \frac{dQ}{dt} = \frac{nqAdx}{dt} = nqA\bar{v}$$

Figure 1.21 *Current (I) flows along a solid semiconductor rod of cross-sectional area* A

We also define the current density $J = \frac{I}{A}$ and hence we obtain Ohm's law or

$$J = nq\bar{v} = \sigma\varepsilon$$

This is known as the *drift current* and it requires the existence of an electric field.

The application of an electric field can also be viewed using energy band diagrams. The well-known electrostatic relationship between electric field and electric potential and energy is given by
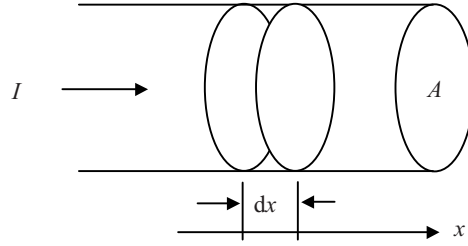
$$\varepsilon(x) = -\frac{dV}{dx} = \frac{1}{q}\frac{dE}{dx} \tag{1.46}$$

which states that an electric field causes a gradient in electric potential V and in addition an electric field causes a gradient in the potential energy E of a charged particle having charge q.

We can represent the conduction and valence bands in an applied electric field by showing the situation where the conduction and valence bands are separated by the energy gap. In Figure 1.7 this occurs at $k = 0$. We introduce spatial dependence by using the x-axis to show the position in the x direction of the semiconductor as in Figure 1.22. If there is no applied field the bands are simply horizontal lines. If a constant electric field is present the energy bands must tilt since from Equation 1.46 there will be a constant gradient in energy and the carriers in each band will experience a force F of magnitude $q\varepsilon$ in the directions shown and will travel so as to lower their potential energies. *The Fermi energy does not tilt*, since the electric field does not change the thermodynamic equilibrium.

We can now describe the flow of electrons. Since $\bar{v} \propto \varepsilon$ we write $\bar{v} = \mu\varepsilon$ where μ is the carrier *mobility*, and we also conclude from Ohm's law that $\sigma = nq\mu$.

In order to confirm the validity of Ohm's law we can start with Newton's law of motion for an electron in an electric field

$$\frac{F}{m^*} = -\frac{q\varepsilon}{m^*} = \frac{d\bar{v}}{dt}$$

The treatment of carrier collisions requires adding the well-known *damping term* $\frac{\bar{v}}{\tau}$ where τ is the scattering time that results in a terminal velocity. This can be pictured by the example of a terminal velocity reached by a ping-pong ball falling in air. We now have

$$\frac{F}{m^*} = -\frac{q\varepsilon}{m^*} = \frac{d\bar{v}}{dt} + \frac{\bar{v}}{\tau} \tag{1.47}$$
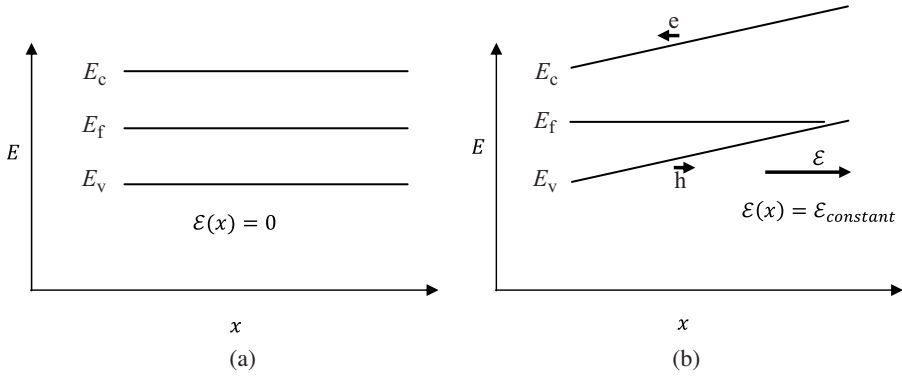
Figure 1.22   *Spatial dependence of energy bands in an intrinsic semiconductor. If there is no electric field (a) the bands are horizontal and electron and hole energies are independent of location within the semiconductor. If an electric field ε is present inside the semiconductor the bands tilt. For an electric field pointing to the right (b), electrons in the conduction band experience a force to the left, which decreases their potential energy. Holes in the valence band experience a force to the right, which decreases their potential energy. This reversed direction for hole energies is described in Figure 1.10*

We can demonstrate the validity of the equation in steady state where $\frac{d\bar{v}}{dt} = 0$ and hence

$$\bar{v} = \frac{q\tau\varepsilon}{m^*}$$

and

$$\mu = \frac{q\tau}{m^*}$$

which are consistent with Ohm's law.

   In addition, we can examine the case where $\varepsilon = 0$. Now from Equation 1.47 we obtain

$$\frac{d\bar{v}}{dt} + \frac{\bar{v}}{\tau} = 0$$

which has solution $\bar{v}(t) = \bar{v}(0)\exp\left(-\frac{t}{\tau}\right)$. Carrier drift velocity will decay upon removal of the electric field with characteristic time constant equal to the scattering time τ.

   In order to consider the contribution of both electrons and holes, we write the total drift current as

$$J_{\text{drift}} = J_{\text{n}} + J_{\text{p}} = q(n\mu_n + p\mu_p)\varepsilon \tag{1.48}$$

where

$$\mu_{\text{n}} = \frac{q\tau}{m_{\text{n}}^*}$$

and

$$\mu_{\text{p}} = \frac{q\tau}{m_{\text{p}}^*}$$

Figure 1.23   *Dependence of drift velocity on electric field for a semiconductor*

Separate mobility values μ_n and μ_p are needed for electrons and holes since they flow in different bands and may have different effective masses m_n^* and m_p^* respectively. The valence band has negative curvature, and Equation 1.16 shows that valence band electrons have negative effective mass; however, to ensure that μ_h is a positive quantity we define the effective mass of holes m_h^* to be a *positive* quantity equal in magnitude to this negative effective mass.

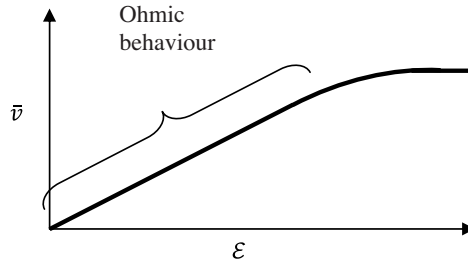The validity of Ohm's law has a limit. If the electric field is large, carrier velocity will usually saturate, as shown in Figure 1.23 and will no longer be linearly proportional to the electric field. This occurs because energetic electrons transfer more energy to lattice vibrations. The magnitude of the electric field that results in saturation effects depends on the semiconductor.

Mobility values for a range of semiconductors are shown in Appendix 2. These are tabulated for intrinsic materials at room temperature; however, defects and impurities as well as higher temperatures have a substantial effect on mobility values since they decrease scattering times. Both undesirable impurities as well as intentionally introduced dopant atoms will cause scattering times and mobility values to decrease.

1.15   Equilibrium and Non-Equilibrium Dynamics

The carrier concentrations we have been discussing until now are *equilibrium* concentrations, and are in thermodynamic equilibrium with the semiconductor material. In equilibrium, both EHP generation and EHP recombination occur simultaneously; however, the net EHP concentration remains constant. We can express this using rate constants defined as follows:

$$G_{th} = \text{thermal EHP generation rate (EHP cm}^{-3}\text{ s}^{-1})$$
$$R = \text{EHP recombination rate (EHP cm}^{-3}\text{ s}^{-1})$$

In equilibrium $G_{th} = R$.

It is easy to cause a *non-equilibrium* condition to exist in a semiconductor. For example, we can illuminate the semiconductor with photons whose energy exceeds its energy gap of the semiconductor. We can also cause electric current to flow through the semiconductor by attaching two or more electrodes to the semiconductor and then connecting them across a voltage source.

In an ideal direct gap semiconductor the value of R depends on carrier concentrations. If, for example, the electron concentration n is doubled, R will double since the probability for an electron to reach a hole has doubled. If both n and p are doubled then R will increase by a factor of four since the hole concentration has also doubled. We can state this mathematically as $R \propto np$. We remove subscripts on carrier concentrations when non-equilibrium conditions are present.

If a steady optical generation rate G_{op} is abruptly added to the thermal generation rate by illumination that begins at time $t = 0$, the total generation rate increases to $G_{th} + G_{op}$. This will cause the EHP generation rate to exceed the recombination rate and carrier concentrations will exceed the equilibrium concentrations and will become time dependent. We shall designate $\delta n(t)$ and $\delta p(t)$ to be the time-dependent carrier concentrations in excess of equilibrium concentrations n_0 and p_0.

Consider a direct gap n-type semiconductor in which $n_0 \gg p_0$. We shall also stipulate that $\delta n(t) \ll n_0$, which states that the excess carrier concentration is small compared to the equilibrium majority carrier concentration. This is referred to as the *low-level injection* condition. We can approximate n to be almost constant and essentially independent of the illumination. Hence the rate of increase of p is determined by the optical generation rate as well as a hole recombination rate term that is linearly proportional to $\delta p(t)$ and we can write the simple differential equation:

$$\frac{d\delta p(t)}{dt} = G_{op} - \frac{\delta p(t)}{\tau_p} \tag{1.49}$$

The solution to this is

$$\delta p(t) = \Delta p \left[1 - \exp\left(\frac{-t}{\tau_p}\right) \right] \tag{1.50}$$

which is sketched in Figure 1.24.

After a time $t \gg \tau_p$, a steady-state value of excess carrier concentration Δp exists. If Equation 1.50 is substituted into Equation 1.49 we obtain

$$\Delta p = G_{op}\tau_p \tag{1.51a}$$

The time constant τ_p for this process is called the *recombination time*. Recombination time is often the same as the *minority carrier lifetime* but should not be confused with the scattering time of Section 1.14, which is generally orders of magnitude shorter than the recombination time. From Equation 1.50, if we know the optical generation rate and
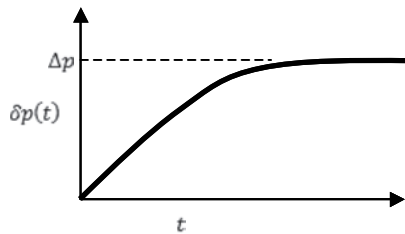

Figure 1.24   *Plot of excess hole concentration as a function of time. A constant optical generation rate starts at* t $= 0$ *and continues indefinitely*

the characteristic recombination time for a semiconductor, we can calculate the steady-state excess minority carrier concentration. There is also a small steady-state change in the majority carrier concentration because $\Delta n = \Delta p$; however, this is neglected in low-level injection since $\Delta n \ll n_0$. Note that $\tau_p \propto \frac{1}{n_0}$.

The same argument can be applied to a p-type semiconductor, and we would obtain the increase in minority carrier concentration thus:

$$\Delta n = G_{\text{op}} \tau_{\text{n}} \tag{1.51b}$$

Electron-hole recombination may also occur via indirect processes in indirect gap semiconductors such as silicon. This is discussed in section 1.19.

1.16   Carrier Diffusion and the Einstein Relation

Free carriers that are produced in a spatially localized part of a semiconductor are able to diffuse and thereby move to other parts of the material. The carrier diffusion process is functionally similar to the diffusion of atoms in solids. At sufficiently high temperatures atomic diffusion occurs, which is described as net atomic motion from a region of higher atomic concentration to a region of lower atomic concentration. This occurs due to random movements of atoms in a concentration gradient. There is no preferred direction to the random movement of the atoms; however, provided the average concentration of atoms is not uniform, the result of random movement is for a net flux of atoms to exist flowing from a more concentrated region to a less concentrated region. *Fick's first law* applies to the diffusion process for atoms as presented in introductory materials science textbooks, and is also applicable to electrons.

As with atomic diffusion, the driving force for carrier diffusion is the gradient in electron concentration. For free electrons diffusing along the *x*-axis, Fick's first law applies and it can be written

$$\phi_{\text{n}}(x) = -D_{\text{n}} \frac{\mathrm{d}n(x)}{\mathrm{d}x} \tag{1.52}$$

where ϕ_{n} is the flux of electrons (number of electrons per unit area per second) flowing along the *x*-axis due to a concentration gradient of electrons. The negative sign in Equation 1.52 indicates that diffusion occurs in the direction of decreasing electron concentration. For holes, Fick's first law becomes

$$\phi_{\text{p}}(x) = -D_{\text{p}} \frac{\mathrm{d}p(x)}{\mathrm{d}x} \tag{1.53}$$

Since the flow of charged particles constitutes an electric current, we can describe *diffusion currents* due to holes or electrons. These are distinct from drift currents described in Section 1.14 because no electric field is involved. Equations 1.52 and 1.53 may be rewritten as currents:

$$J_{\text{n}}(x)_{\text{diffusion}} = q\,D_{\text{n}} \frac{\mathrm{d}n(x)}{\mathrm{d}x} \tag{1.54a}$$

$$J_{\text{p}}(x)_{\text{diffusion}} = -q\,D_{\text{p}} \frac{\mathrm{d}p(x)}{\mathrm{d}x} \tag{1.54b}$$

Note that there is no negative sign in the case of electrons because electrons carry a negative charge that cancels out the negative sign in Fick's first law.

An interesting situation occurs when both diffusion and drift currents flow. An electric field is present as well as a carrier concentration gradient. The total current densities from Equations 1.48 and 1.54 become

$$J_n(x)_{\text{drift+diffusion}} = q\mu_n n(x)\varepsilon(x) + qD_n\frac{dn(x)}{dx} \tag{1.55a}$$

and

$$J_p(x)_{\text{drift+diffusion}} = q\mu_p p(x)\varepsilon(x) - qD_p\frac{dp(x)}{dx} \tag{1.55b}$$

In semiconductor diodes, both drift and diffusion occur and it is important to become familiar with the situation where drift and diffusion currents coexist in the same part of the semiconductor.

One way to establish an electric field is to have a gradient in doping level by spatially varying the doping concentration in the semiconductor. Consider the example in Figure 1.25. The dopant concentration varies across a semiconductor sample that is in thermal equilibrium. On the left side, the semiconductor is undoped, and an acceptor dopant gradually increases in concentration from left to right. This causes the Fermi energy to occupy lower positions in the energy gap until it is close to the valence band on the right
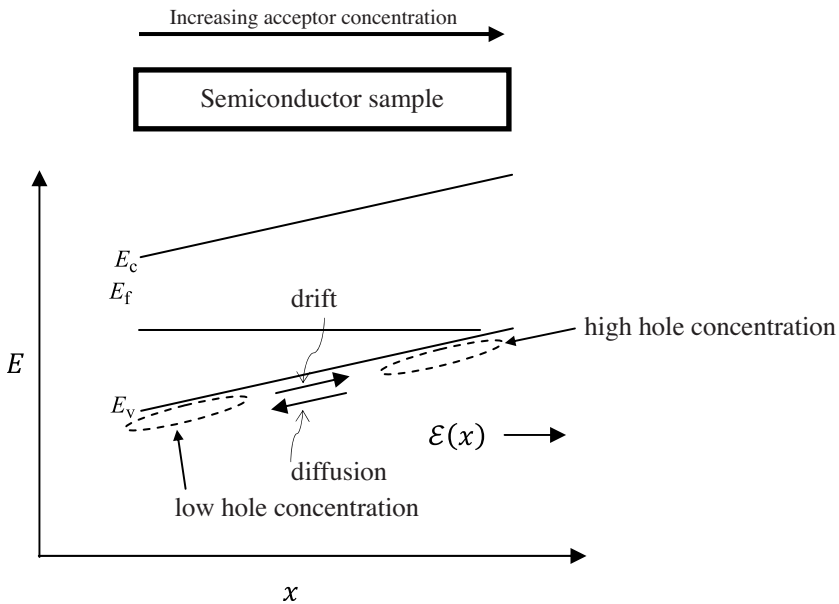

Figure 1.25   *The energy bands will tilt due to a doping gradient. Acceptor concentration increases from left to right in a semiconductor sample. This causes a built-in electric field, and the hole concentration increases from left to right. The field causes hole* drift *from left to right, and there is also hole* diffusion *from right to left due to the concentration gradient*

side of the sample. The Fermi energy does not tilt because it is a thermodynamic quantity and the sample is in equilibrium; however, the valence and conduction bands do tilt as shown.

There will now be a high concentration of holes in the valence band on the right side of the semiconductor, which decreases to a low hole concentration on the left side. Hole diffusion will therefore occur in the negative-*x* direction. At the same time, the tilting of the energy bands means that an electric field is present in the sample. This is known as a *built-in electric field* since it is caused by a spatial concentration variation within the semiconductor material rather than by the application of an applied voltage. The built-in field causes a hole drift current to flow in the positive-*x* direction. Since the semiconductor is in equilibrium, these two hole currents cancel out and *the net hole current flow will be zero.* A similar argument can be made for the electrons in the conduction band and the net electron current will also be zero.

A useful relationship between mobility and diffusivity can now be derived. Since the net current flow illustrated in Figure 1.25 is zero in equilibrium we can write for hole current

$$J_p(x)_{\text{drift+diffusion}} = q\mu_p p(x)\varepsilon(x) - qD_p\frac{dp(x)}{dx} = 0 \tag{1.56}$$

From Equations 1.46 and 1.56 we have

$$q\mu_p p(x)\frac{1}{q}\frac{dE}{dx} - qD_p\frac{dp(x)}{dx} = 0$$

We now calculate $p(x)$. Since the valence band energy E_v is now a function of x we can rewrite Equation 1.35 as

$$p(x) = N_v \exp\left(\frac{-(E_f - E_v(x))}{kT}\right) \tag{1.57}$$

and we obtain

$$\mu_p N_v \exp\left(\frac{-(E_f - E_v(x))}{kT}\right)\frac{dE_v(x)}{dx} - qD_p\frac{N_v}{kT}\exp\left(\frac{-(E_f - E_v(x))}{kT}\right)\frac{dE_v(x)}{dx} = 0$$

which simplifies to

$$\frac{D_p}{\mu_p} = \frac{kT}{q} \tag{1.58a}$$

and a similar derivation may be applied to electrons yielding

$$\frac{D_n}{\mu_n} = \frac{kT}{q} \tag{1.58b}$$

Equation 1.58 is known as the *Einstein relation.* At a given temperature this tells us that mobility and diffusivity are related by a constant factor, which is not unexpected since both

quantities express the degree of ease with which carriers move in a semiconductor under a driving force.

1.17 Quasi-Fermi Energies

If a semiconductor is influenced by incident photons or an applied electric current the semiconductor is no longer in equilibrium. This means that we cannot use Equations 1.34 and 1.35 to determine carrier concentrations. In addition Fermi energy E_f is no longer a meaningful quantity since it was defined for a semiconductor in equilibrium in Section 1.8 and the Fermi–Dirac distribution function of Figure 1.9 is also based on equilibrium conditions.

For convenience, we define two new quantities, F_n and F_p, known as the *quasi-Fermi energy for electrons* and the *quasi-Fermi energy for holes*, respectively. The quantities may be used even if a semiconductor is not in equilibrium and there are excess carriers. F_n and F_p are defined for a semiconductor with excess carriers from the following equations:

$$n = N_c \exp\left(\frac{-(E_c - F_n)}{kT}\right) \tag{1.59a}$$

and

$$p = N_v \exp\left(\frac{-(F_p - E_v)}{kT}\right) \tag{1.59b}$$

Note the similarity between Equation 1.59 and Equations 1.34 and 1.35. Subscripts for n and p are absent in Equation 1.59 because these carrier concentrations are not necessarily equilibrium values. It follows that if the semiconductor is in equilibrium, the electron and hole quasi-Fermi energies become equal to each other and identical to the Fermi energy. The electron and hole quasi-Fermi levels in an n-type semiconductor will behave very differently upon excess carrier generation, as shown in Figure 1.26. This is examined in Example 1.4.
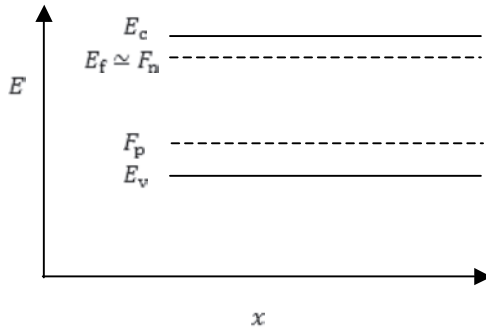

Figure 1.26 *The quasi-Fermi levels F_n and F_p for an n-type semiconductor with excess carriers generated by illumination. Note the large change in F_p due to illumination and note that F_n is almost the same as the value of E_f before illumination*

Example 1.4

An n-type silicon sample has a donor concentration of 1×10^{17} cm^{-3}. In Example 1.3a we obtained $n_0 = 1 \times 10^{17}$ cm^{-3}, $p_0 = 2.25 \times 10^3$ cm^{-3} and $E_c - E_f = 0.15$ eV. We now illuminate this sample and introduce a uniform electron-hole pair generation rate of $G_{op} = 5 \times 10^{20}$ cm^{-3} s^{-1}. Assume a carrier lifetime of 2×10^{-6} s.

(a) Calculate the resulting electron and hole concentrations.
(b) Calculate the quasi-Fermi energy levels.

Solution

(a)

$$\Delta p = G_{op}\tau_p = 5 \times 10^{20} \text{ cm}^{-3} \text{ s}^{-1} \times 2 \times 10^{-6} \text{ s} = 1 \times 10^{15} \text{ cm}^{-3}$$

and

$$\Delta n = G_{op}\tau_n = 5 \times 10^{20} \text{ cm}^{-3} \text{ s}^{-1} \times 2 \times 10^{-6} \text{ s} = 1 \times 10^{15} \text{ cm}^{-3}$$

Hence

$$p = p_0 + \Delta p = 2.25 \times 10^3 \text{ cm}^{-3} + 1 \times 10^{15} \text{ cm}^{-3} \cong 1 \times 10^{15} \text{ cm}^{-3}$$

and

$$n = n_0 + \Delta n = 1 \times 10^{17} \text{ cm}^{-3} + 1 \times 10^{15} \text{ cm}^{-3} = 1.01 \times 10^{17} \text{ cm}^{-3}$$

Therefore the carrier concentrations may be strongly affected by the illumination: the hole concentration increases by approximately 12 orders of magnitude from a very small minority carrier concentration to a much larger value dominated by the excess hole concentration.

The electron concentration, however, only increases slightly (by 1%) due to the illumination since it is a majority carrier. This is therefore an example of low-level injection since the majority carrier concentration is almost unchanged.

(b) The quasi-Fermi level for holes may be found from:

$$p = N_v \exp\left(\frac{-(F_p - E_v)}{kT}\right)$$
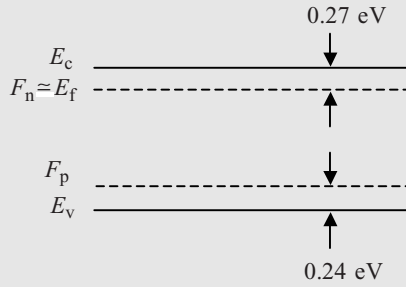
Solving for $F_p - E_v$ we obtain

$$F_p - E_v = kT \ln\left(\frac{N_v}{p}\right) = 0.026 \text{ eV } \ln\left(\frac{1.06 \times 10^{19} \text{ cm}^{-3}}{1 \times 10^{15} \text{ cm}^{-3}}\right) = 0.24 \text{ eV}$$

The quasi-Fermi level for electrons may be found from

$$n = N_c \exp\left(\frac{-(E_c - F_n)}{kT}\right)$$

Solving for $E_c - F_n$ we obtain

$$E_c - F_n = kT \ln \left(\frac{N_c}{n} \right) = 0.026 \text{ eV } \ln \left(\frac{2.84 \times 10^{19} \text{ cm}^{-3}}{1.01 \times 10^{15} \text{ cm}^{-3}} \right) = 0.27 \text{ eV}$$

0.27 eV

E_c ——————————————

$F_n \simeq E_f$ -

F_p -

E_v ——————————————

0.24 eV

Note that with illumination F_n is almost identical to the original value of E_f but F_p moves significantly lower. This is a consequence of the large excess carrier concentration compared to the equilibrium hole concentration. If a similar problem were solved for a p-type semiconductor then with illumination F_n would move significantly higher and F_p would remain almost identical to the original value of E_f.

The separation between F_n and F_p is a measure of the product of excess carrier concentrations divided by the product of equilibrium carrier concentrations since from Equations 1.59 and 1.41 we have

$$np = N_c N_v \exp \left(\frac{F_n - F_p - E_g}{kT} \right) = n_0 p_0 \exp \left(\frac{F_n - F_p}{kT} \right) \qquad (1.60a)$$

and hence

$$F_n - F_p = kT \ln \left(\frac{np}{n_0 p_0} \right) \qquad (1.60b)$$

1.18   The Diffusion Equation

We have introduced carrier recombination as well as carrier diffusion separately; however, carriers in semiconductors routinely undergo both diffusion and recombination simultaneously.

In order to describe this, consider a long semiconductor bar or rod in Figure 1.27 in which excess holes are generated at $x = 0$ causing an excess of holes Δp to be maintained at $x = 0$. The excess hole concentration drops off to approach an equilibrium concentration at the other end of the rod. The excess holes will diffuse to the right and if an electric field is present there will also be a drift current component. Some of these holes recombine
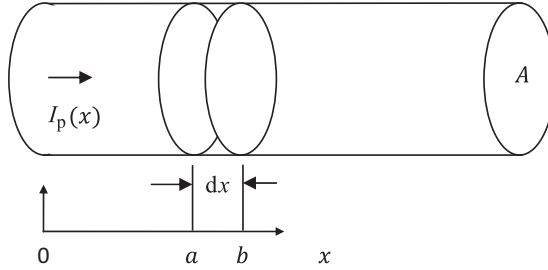
Figure 1.27 *A solid semiconductor rod of cross-sectional area A has a hole current I$_p$(x) flowing in the positive x direction. Due to recombination the hole current is dependent on x. At surfaces x = a and x = b, I$_p$(x) changes due to the recombination that occurs between these of the rod surfaces within volume Adx*

with electrons during this process. We can consider a slice of width dx as shown in Figure 1.27. The hole current $I_p (x = a)$ will be higher than the hole current $I_p (x = b)$ due to the rate of recombination of holes in volume Adx between $x = a$ and $x = b$. Volume Adx contains $Adx\delta p(x)$ excess holes. Since $\frac{\delta p(x)}{\tau_p}$ is the recombination rate this may be expressed mathematically:

$$I_p (x = a) - I_p (x = b) = -qAdx\frac{\delta p(x)}{\tau_p} \tag{1.61}$$

This may be rewritten as

$$\frac{I_p (x = a) - I_p (x = b)}{dx} = -qA\frac{\delta p(x)}{\tau_p}$$

or

$$\frac{dI_p(x)}{dx} = -qA\frac{\delta p(x)}{\tau_p}$$

In terms of current density we have

$$\frac{dJ_p(x)}{dx} = -q\frac{\delta p(x)}{\tau_p} \tag{1.62a}$$

and applying the same procedure to electrons we obtain

$$\frac{dJ_n(x)}{dx} = -q\frac{\delta n(x)}{\tau_n} \tag{1.62b}$$

If the current is entirely due to the *diffusion* of carriers, we rewrite the expression for diffusion current from Equation 1.54 for excess carriers $\delta p(x)$:

$$J_p(x)_{\text{diffusion}} = -qD_p\frac{d\delta p(x)}{dx} \tag{1.63}$$

Substituting this into Equation 1.62a we obtain

$$\frac{d^2 \delta p(x)}{dx^2} = \frac{\delta p(x)}{D_p \tau_p} \tag{1.64a}$$

This is known as the *steady state diffusion equation for holes*, and the corresponding equation for electrons is:

$$\frac{d^2 \delta n(x)}{dx^2} = \frac{\delta n(x)}{D_n \tau_n} \tag{1.64b}$$

The general solution to Equation 1.64a is

$$\delta p(x) = A \exp\left(\frac{-x}{\sqrt{D_p \tau_p}}\right) + B \exp\left(\frac{x}{\sqrt{D_p \tau_p}}\right) \tag{1.65a}$$

However, considering our boundary conditions, the function must decay to zero for large values of x and therefore $B = 0$ yielding

$$\delta p(x) = \Delta p \exp\left(\frac{-x}{\sqrt{D_p \tau_p}}\right)$$

which may be written

$$\delta p(x) = \Delta p \exp\left(\frac{-x}{L_p}\right) \tag{1.65b}$$

where

$$L_p = \sqrt{D_p \tau_p}$$

is known as the *diffusion length*. The latter determines the position on the x-axis where carrier concentrations are reduced by a factor of e, as shown in Figure 1.28.
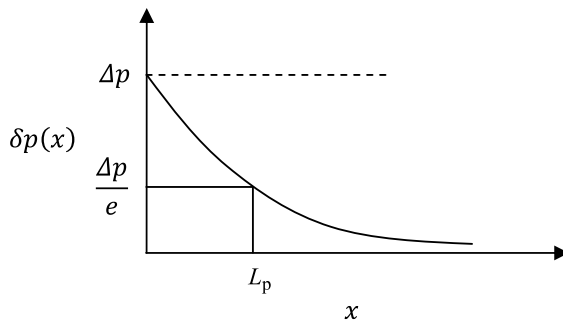

Figure 1.28 *Plot of excess hole concentration in a semiconductor as a function of x in a semiconductor rod where both diffusion and recombination occur simultaneously. The decay of the concentration is characterized by a diffusion length L_p*
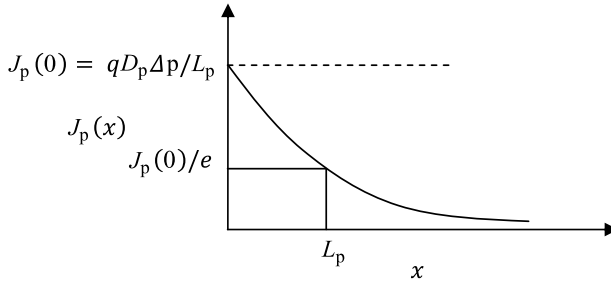
Figure 1.29 *Hole current density as a function of* x *for a semiconductor rod with excess carriers generated at* x = 0

The hole current density at any point x may be determined by substituting Equation 1.65b into Equation 1.63, and we obtain

$$J_p(x)_{\text{diffusion}} = q \frac{D_p}{L_p} \Delta p \, \exp\left(\frac{-x}{L_p}\right)$$

This may also be written

$$J_p(x)_{\text{diffusion}} = q \frac{D_p}{L_p} \delta p(x)$$

which shows that both current density and hole concentration have the same exponential form. Figure 1.29 plots current density $J_p(x)_{\text{diffusion}}$ as a function of x. Both diffusion and recombination occur simultaneously, which lowers the diffusion current exponentially as x increases. This is reasonable since the number of holes that have not recombined drops exponentially with x and therefore L_p also represents the position on the x-axis where current density $J_p(x)$ is reduced by a factor of e.

1.19   Traps and Carrier Lifetimes

Carrier lifetimes in many semiconductors including indirect gap silicon are really determined by trapping processes instead of ideal direct gap electron-hole recombination. *Traps* are impurity atoms or native point defects such as vacancies, dislocations or grain boundaries. There are also *surface traps* due to the defects that inherently occur at semiconductor surfaces, and *interface traps* that form at a boundary between two different material regions in a semiconductor device.

In all cases, traps are physical defects capable of trapping conduction band electrons and/or valence band holes and affecting carrier concentrations, carrier flow and recombination times. This is different from a normal electron-hole pair recombination process because a specific defect is involved. After being trapped a carrier may again be released to the band it originated from or it may subsequently recombine at the trap with a carrier of the opposite sign that also gets attracted to the same trap. This is known as *trap-assisted carrier recombination* and is one of the most important phenomena that limits the performance of both direct and indirect gap semiconductors used for solar cells and LEDs.

Intentionally introduced n-type and p-type dopants actually are traps, and are referred to as *shallow traps* because they are only separated from either a conduction band or a valence band by a small energy difference that may be overcome by thermal energy. This means that the trap is easy to ionize and the carrier is very likely to be released from the trap. Also since dopants are normally ionized, they do not trap carriers of the opposite sign. For example, phosphorus in silicon is an n-type dopant. Once it becomes a positive ion after donating an electron to the conduction band it has a small but finite probability of recapturing a conduction band electron. Since it spends almost all its time as a positive ion, it has essentially no chance of capturing a hole, which is repelled by the positive charge, and therefore shallow traps generally do not cause trap-assisted carrier recombination.

The traps that we must pay careful attention to are *deep traps*, which exist near the middle of the bandgap. These traps are highly effective at promoting electron-hole pair recombination events. Since their energy levels are well separated from band edges, carriers that are trapped are not easily released. Imagine a deep trap that captures a conduction band electron and is then negatively charged. In this state the negatively charged trap cannot readily release its trapped electron and may therefore attract a positive charge and act as an effective hole trap. Once the hole is trapped it recombines with the trapped electron and the trap is effectively emptied and is again available to trap another conduction band electron. In this manner, traps become a new conduit for electrons and holes to recombine. If the deep trap density is high the average trap-assisted recombination rate is high.

The analysis of deep trap behaviour requires that we know the probabilities of the trap being filled or empty. This may be understood in equilibrium conditions by knowing the trap energy level and comparing it to the Fermi energy level. If the trap energy level E_t is above E_f then the trap is more likely to be empty than full. If E_t is below E_f it is more likely to be filled. The terms 'filled' and 'empty' refer specifically to electrons because the Fermi–Dirac function describes the probability that an electron fills a specific energy level. If we wish to describe the probabilities for a trap to be occupied by holes, we must subtract these probabilities from 1.

We will simplify the treatment of traps by focusing on a very specific situation. Consider a trap *at the Fermi energy and near mid-gap* in a semiconductor. Both E_t and E_f will be at approximately the middle of the energy gap. The ionization energy of the trap for either a trapped electron or a trapped hole is approximately $\frac{E_g}{2}$, as illustrated in Figure 1.30. The probability that the trap is empty or filled is 50% since it is at the Fermi energy. The captured electron may be re-released back to the conduction band, or it may be annihilated by a hole at the trap.

There is a simple argument for assuming the trap is likely to exist at the Fermi level and near mid-gap. At the surface of a semiconductor, approximately half the bonds normally formed will be broken or incomplete. This gives rise to electrons that are only held by half the atoms that they usually are associated with in the interior of a perfect crystal. These
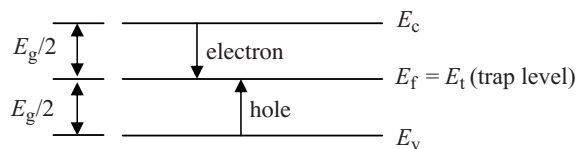

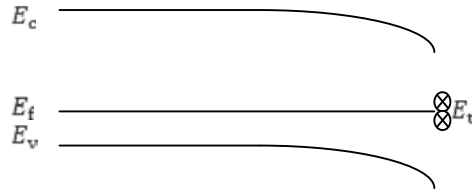**Figure 1.30**   *A trap level at the Fermi energy near mid-gap*

Figure 1.31   *Surface traps at the surface of a p-type semiconductor comprise electrons held in dangling bonds. The energy needed to release these electrons is approximately $\frac{E_g}{2}$. Since there are large numbers of dangling bond states, some being occupied and some not being occupied by electrons, the Fermi energy becomes pinned at this energy*

dangling bonds therefore comprise electrons that are likely to lie at approximately mid-gap because the energy required to excite them into the conduction band is only approximately half as large as the energy E_g required to remove an electron from the complete covalent bond of the relevant perfect crystal.

Now consider a large number of such dangling bonds at a semiconductor surface. Some of these dangling bonds will have lost electrons and some of them will not. Since the Fermi level exists between the highest filled states and the lowest empty state, the Fermi level tends to fall right onto the energy level range of these traps. The Fermi energy gets *pinned* to this trap energy at $\cong \frac{E_g}{2}$. Figure 1.31 shows the pinning of a Fermi level due to surface traps in a p-type semiconductor. Notice that at the semiconductor surface the surface traps determine the position of the Fermi energy rather than the doping level. An electric field is established in the semiconductor normal to the surface and *band bending* occurs as shown.

If the semiconductor had been n-type instead of p-type, then the same reasoning would still pin the Fermi energy to mid-gap; however, the band bending would occur in the opposite direction and the resulting electric field would point in the opposite direction, as illustrated in Figure 1.32.

Since traps are often formed from defects other than free surfaces that also involve incomplete bonding like a vacancy, a dislocation line, a grain boundary or an interface between two layers, this simple picture is very useful and will be used in the context of the p-n junction to explain recombination processes in subsequent sections of this book.

There is a velocity associated with excess minority carriers at a semiconductor surface or an interface between a semiconductor and another material. For example in Figure 1.31, which shows a p-type semiconductor, assume that there are excess electrons $\delta n(x)$ in the conduction band. The conduction band excess electron concentration decreases towards the
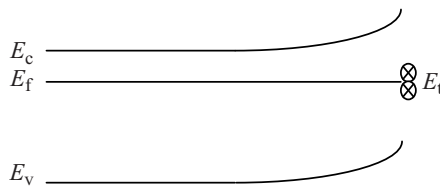

Figure 1.32   *Surface traps at the surface of an n-type semiconductor causing the Fermi level to be trapped at approximately mid-gap. An electric field opposite in direction to that of Figure 1.28 is formed in the semiconductor*

surface. This causes electrons to diffuse towards the surface, where they recombine. The diffusion of electrons towards the surface can be expressed as an electron flux

$$\varphi = -D_n \frac{d\delta n}{dx}$$

Flux has units of particles per unit area per unit time. Since we can equivalently define a flux as the product of concentration and velocity, we can write

$$\varphi = -D_n \frac{d\delta n}{dx} = S_n \delta n \tag{1.66a}$$

where S_n is the *surface recombination velocity* of the electrons, and we evaluate δn and $\frac{d\delta n}{dx}$ at the semiconductor surface. Note that if $S_n = 0$, we can conclude that $\frac{d\delta n}{dx}$ is zero, there is no band bending and surface states do not form a charged layer. Conversely if $S_n \to \infty$ then $\delta n = 0$ at the surface, which implies that carriers very rapidly recombine at the surface. A similar situation exists at the surface or at an interface of an n-type semiconductor in which holes may recombine and we obtain

$$\varphi = -D_p \frac{d\delta p}{dx} = S_p \delta p \tag{1.66b}$$

In both Equations 1.66a and 1.66b we assume that the excess carrier concentrations are much larger than the equilibrium minority carrier concentrations.

1.20   Alloy Semiconductors

An important variation in semiconductor compositions involves the use of partial substitutions of elements to modify composition. One example is the partial substitution of germanium in silicon that results in a range of new semiconductors of composition $Si_{1-x}Ge_x$, which are known as *alloy semiconductors*. The germanium atoms randomly occupy lattice sites normally occupied by silicon atoms, and the crystal structure of silicon is maintained. Note that Si and Ge are both in the group IV column of the periodic table and therefore have chemical similarities in terms of valence electrons and types of bonding. This means that provided no additional dopant impurities are introduced into the alloy semiconductor, alloy material with characteristics of an intrinsic semiconductor can be achieved. Of interest in semiconductor devices is the opportunity to modify the optical and electrical properties of the semiconductor. Since germanium has a smaller bandgap than silicon, adding germanium decreases bandgap as x increases. In addition, the average lattice constant of the new compound will increase since germanium is a larger atom than silicon. Since both germanium and silicon have the same diamond crystal structure, the available range of x is from 0 to 1 and the indirect bandgaps of the alloy compositions $Si_{1-x}Ge_x$ therefore range between 1.11 eV and 0.067 eV as *x* varies from 0 to 1 respectively.

Of more relevance to p-n junctions for solar cells and LEDs, alloy semiconductors may also be formed from compound semiconductors. For example $Ga_{1-x}In_xN$ is a *ternary*, or three-component, alloy semiconductor in which a fraction of the gallium atoms in wurtzite GaN is replaced by indium atoms. The indium atoms randomly occupy the crystalline sites in GaN that are normally occupied by gallium atoms. Since In and Ga are both group

III elements, the substitution does not act as either an acceptor or a donor. The direct bandgap decreases as x increases. For $x = 0$, $E_g = 3.4$ eV and for $x = 1$, $E_g = 0.77$ eV.

There are many other III-V alloy semiconductors. In $Ga_{1-x}Al_xAs$ alloys the bandgap varies from 1.43 to 2.16 eV as x goes from 0 to 1. In this system, however, the bandgap is direct in the case of GaAs, but indirect in the case of AlAs. There is a transition from direct to indirect bandgap at $x \cong 0.38$. We can understand this transition if we consider the two conduction band minima in GaAs shown in Figure 1.16c. One minimum forms a direct energy gap with the highest energy levels in the valence band; however, the second minimum forms an indirect gap. When $x \cong 0.38$ these two minima are at the same energy level. For $x < 0.38$ the alloy has a direct gap because the global conduction band minimum forms the direct gap. For $x > 0.38$ the global conduction band minimum is the minimum that forms the indirect gap.

A number of III-V alloy systems are illustrated in Figure 1.33a. An additional set of III-V nitride semiconductors is shown in Figure 1.33b, and a set of II-VI semiconductors is included in Figure 1.33c.
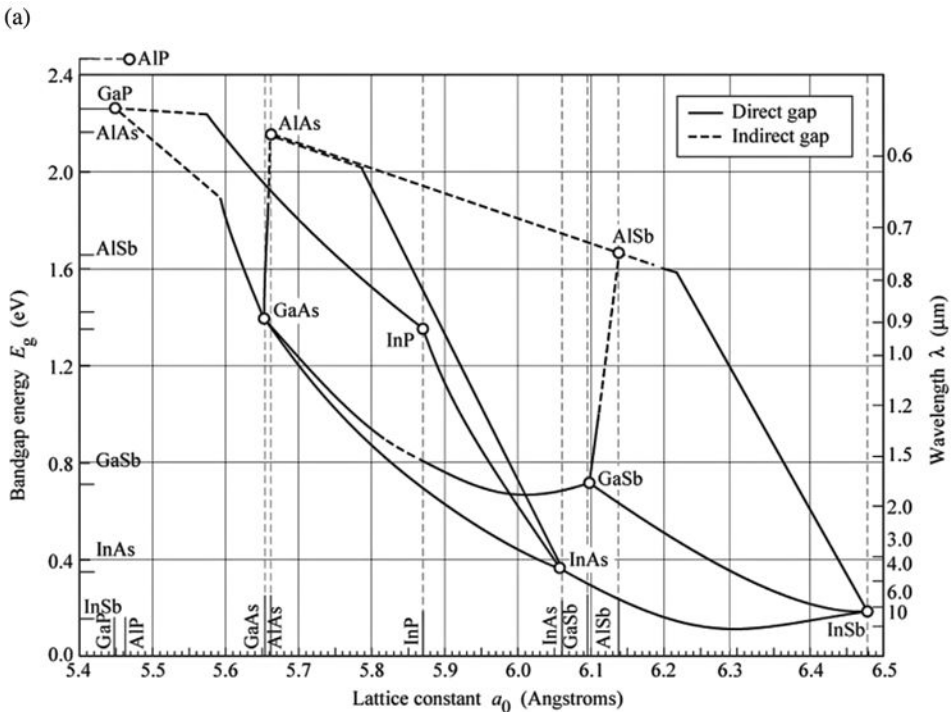

Figure 1.33 *Bandgap versus lattice constant for (a) phosphide, arsenide and antimonide III-V semiconductors; (b) nitride and other III-V semiconductors neglecting bowing (see Figure 5.20); (c) sulphide, selenide and telluride II-VI semiconductors and phosphide, arsenide and antimonide III-V semiconductors. (a–b) Reprinted from E. Fred Schubert, Light-Emitting Diodes, 2e ISBN 978-0-521-86538-8. Copyright (2006) with permission from E. Fred Schubert. (c) Adapted from http://www.tf.uni-kiel.de/matwis/amat/semitech_en/ Copyright Prof. Dr. Helmut Föll, Technische Fakultät, Universität Kiel*

(b)
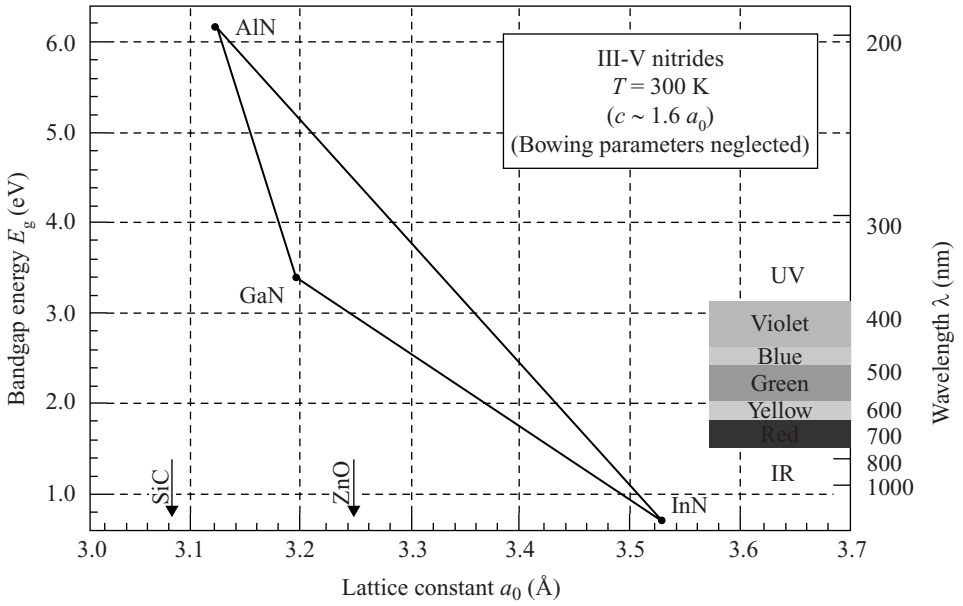
(c)

Figure 1.33   (Continued)

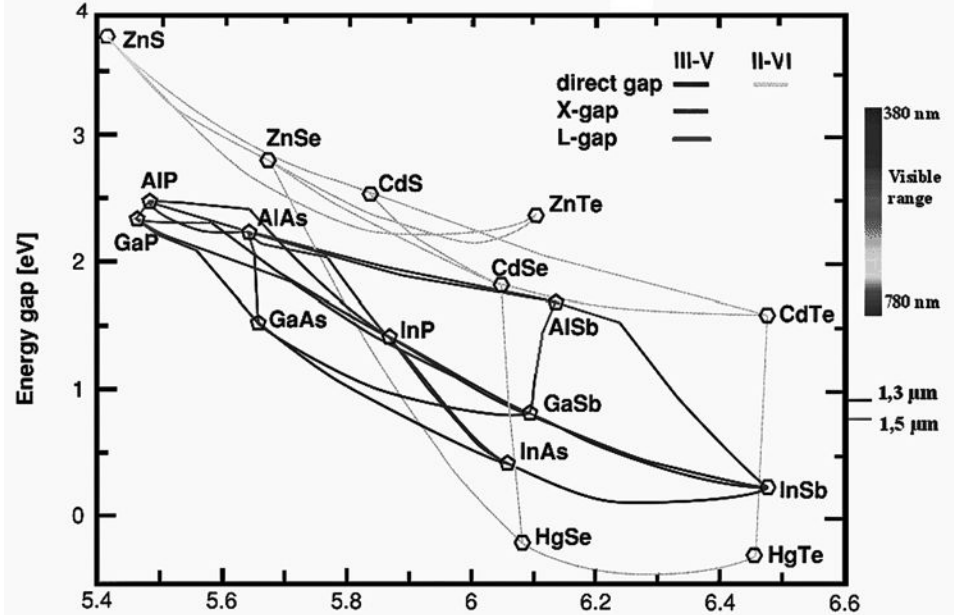To cover ranges of composition in the III-V alloy system, *quaternary* alloys may be formed such as $(Al_xGa_{1-x})_yIn_{1-y}P$. The use of ternary and quaternary semiconductor alloys in solar cells and LEDs will be discussed in Chapters 4 and 5.

1.21   Summary

1.1. The band theory of solids permits an understanding of electrical and optical properties including electrical conductivity in metals and semiconductors, optical absorption and luminescence, and properties of junctions and surfaces of semiconductors and metals.

1.2. The band theory of solids requires the use of wavefunctions to describe the spatial distribution, energy and momentum of electrons. Schrödinger's equation may be solved to obtain these wavefunctions by employing a suitable potential energy function in this equation.

1.3. In the Kronig–Penney model a periodic potential leads to energy bands, energy gaps and Brillouin zones by solving Schrödinger's equation for electrons in a periodic potential. The size of the energy gaps increases as the amplitude of the periodic potential increases. As the ionic character of the bonding in the semiconductor increases the energy gap increases. As the size of the atoms decreases the energy gap increases.

1.4. The Bragg model identifies Brillouin zone boundaries as satisfying the Bragg condition for strong reflection. This condition is $2a = n\lambda$. The reduced zone scheme, which shows only the first Brillouin zone, simplifies the representation of energy bands and energy gaps.

1.5. The effective mass m^* is used to quantify electron behaviour in response to an applied force. The effective mass depends on the band curvature. Effective mass is constant if the band shape can be approximated as parabolic.

1.6. The number of states in a band n can be determined based on the number of unit cells N in the semiconductor sample. The result $n = 2N$ is obtained for a one-dimensional, two-dimensional or three-dimensional case.

1.7. The filling of bands in semiconductors and insulators is such that the highest filled band is full and the lowest empty band is empty at low temperatures. In metals the highest filled band is only partly filled. Semiconductors have smaller bandgaps ($E_g = 0$ to $4\,eV$) than insulators ($E_g > 4\,eV$).

1.8. The Fermi energy E_f is defined as the energy level at which an electron state has a 50% probability of occupancy at temperatures above $0\,K$. A hole can be created when an electron from the valence band is excited to the conduction band. The hole can move independently from the electron.

1.9. Carrier concentration in an energy band is determined by (i) finding the probability of occupancy of the states in a band using the Fermi–Dirac distribution function $F(E)$, and (ii) finding the density of states function $D(E)$ for an energy band. Then the integral over the energy range of the band of the product of $D(E)$ and $F(E)$ will determine the number of carriers in the band. In the conduction band the equilibrium electron concentration is n_0 and in the valence band the equilibrium hole concentration is p_0. The product $n_0 p_0$ is a constant that is independent of the Fermi energy.

1.10. A range of semiconductor materials includes group IV semiconductors, group III-V semiconductors and group II-VI semiconductors as listed in order of increasing ionic character. Bandgap energies decrease for larger atoms that are lower down on the periodic table.

1.11. Most important semiconductor crystals have lattices that are diamond, zincblende or hexagonal structures. These structures lead to complex band diagrams. The band shapes in E versus k plots depend on crystallographic directions. In addition there are sub-bands in the valence band that correspond to distinct hole effective masses. Bandgaps may be direct or indirect. Conduction bands generally exhibit two minima where one minimum corresponds to a direct gap transition and one minimum corresponds to an indirect gap transition.

1.12. Photon momentum is very small and direct gap transitions in semiconductors are favourable for photon creation and absorption. In indirect gap transitions the involvement of lattice vibrations or phonons is required. The absorption coefficient α is higher for direct gap semiconductors and lower for indirect gap semiconductors for photons above E_g in energy. Whereas indirect gap silicon has an effective absorption depth of $\cong 100\,\mu m$ for sunlight the corresponding absorption depth in GaAs is only $\cong 1\,\mu m$.

1.13. Pure semiconductors are known as intrinsic semiconductors. The incorporation of low levels of impurity atoms in a semiconductor leads to extrinsic semiconductors, in which the electron concentration n_0 and hole concentration p_0 are controlled by the impurity type and concentration. Donor impurities donate electrons to the conduction band in n-type semiconductors, and acceptor impurities donate holes to the valence band in p-type semiconductors. New shallow energy levels arise within the energy gap, which are called donor and acceptor levels. Carrier concentrations are temperature dependent; however, over a wide intermediate temperature range carrier concentrations are relatively constant as a function of temperature. Minority carriers refer to the carriers having a low concentration in a specific semiconductor region, and majority carriers refer to the carriers having a significantly higher concentration in the same region.

1.14. Carriers move through semiconductors in an electric field ε by a drift process, which is characterized by a drift velocity $\bar{v}$ and mobility μ. Drift current density is given by $J = \sigma\varepsilon$, which is an expression of Ohm's law. The understanding of Ohm's law is based on the concept of a terminal velocity due to scattering events having a characteristic scattering time τ, which depends on impurities, defects and temperature. At high electric fields drift velocity will eventually saturate.

1.15. Carrier concentrations are not necessarily at equilibrium levels. Photons or applied electric fields can give rise to non-equilibrium excess carrier concentrations, which will return to equilibrium concentrations once equilibrium conditions are restored. Electron-hole pair (EHP) generation and recombination processes G_{th} or G_{op} and R define the resulting rate of generation and recombination. Under equilibrium conditions $G_{th} = R$ where $R \propto np$. The minority carrier lifetime time constants τ_n or τ_p characterize the recombination times of minority carriers.

1.16. Carriers diffuse in semiconductors due to a concentration gradient. The diffusion coefficients D_n and D_p determine the diffusion current J_n and J_p respectively. The net current flow must include both drift and diffusion current. In equilibrium the net

current is zero; however, drift and diffusion currents may be non-zero. The Einstein relation is derived from the requirements for equilibrium conditions and allows D_n and D_p to be derived from μ_n and μ_p.

1.17. In non-equilibrium conditions the Fermi energy is not defined; however, quasi-Fermi energies F_n and F_p may be defined to characterize changes in carrier concentrations due to excess carrier generation.

1.18. Combining the concepts of carrier recombination and carrier diffusion the diffusion equation leads to the calculation of diffusion length $L_n = \sqrt{D_n \tau_n}$ and $L_p = \sqrt{D_p \tau_p}$ and shows that carrier concentration decays exponentially as a function of distance from a region of excess carrier generation.

1.19. Traps can have a large effect on carrier lifetimes. The most important traps are deep traps that are at or near mid-gap. A high density of such traps occurs at semiconductor interfaces and defects. This is due to dangling bonds at surfaces and defects. Fermi level pinning occurs at or near midgap due to dangling bonds. Surface recombination velocity is a measure of the rate of recombination at semiconductor surfaces.

1.20. The bandgap and direct/indirect nature of a semiconductor can be altered by alloying. Alloying may also change the lattice constant. Industrially important alloy semiconductors exist composed from group IV elements, group III-V elements and group II-VI elements. Ternary and quaternary compound semiconductors are important for solar cells and LEDs.

Suggestions for Further Reading

Ashcroft NW and Mermin ND. *Solid State Physics*. Holt, Rinehart and Winston, 1976.

Kittel C. *Introduction to Solid State Physics*, 8th edn. John Wiley and Sons, 2005.

Eisberg R and Resnick R. *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd edn. John Wiley and Sons, 1985.

Neamen DA. *Semiconductor Physics and Devices*, 3rd edn. McGraw Hill, 2003.

Solymar L and Walsh D. *Electrical Properties of Materials*, 7th edn. Oxford University Press, 2004.

Problems

1.1 Derive the following relationships in the Kronig–Penney model:
   (a) Obtain Equation 1.7c from Equations 1.4a, 1.4b, 1.7a and 1.7b.
   (b) Obtain Equation 1.8 from Equation 1.7c. Take the limit $b \to 0$ and $V_0 \to \infty$ such that bV_0 is constant.
   Use the definition

$$P = \frac{Q^2 ba}{2}$$

1.2 Using the Kronig–Penney model in Equation 1.8:
   (a) If $P = 2$ and $a = 2.5$ Å find the energy range for the two lowest energy bands and the energy gap between them.
   (b) Repeat (a) if $P = 0.2$.
   (c) Demonstrate that the energy bands reduce to free electron states as $P \to 0$.

1.3 In Section 1.6 we showed that the number of states in an energy band is $n = 2N$ for a one-dimensional semiconductor. Show that the number of states in an energy band in a three-dimensional semiconductor is still $n = 2N$ where N is the number of unit cells in the three-dimensional semiconductor.
   Hint: Consider a semiconductor in the form of a rectangular box having N_x, N_y and N_z unit cells along the x, y and z axes. Assume an infinite walled box with $V = 0$ inside the box. Use the allowed energy values for an electron in Equation 1.28. Include spin.

1.4 A rectangular semiconductor crystal has dimensions $2 \times 2 \times 1$ mm. The unit cell is cubic and has edge length of 2 Å. Find the number of states in one band of this semiconductor.

1.5 A rectangular silicon semiconductor bar of length 12 cm and cross-section 1×5 mm is uniformly doped n-type with concentration $N_d = 5 \times 10^{16}$ cm^{-3}.
   (a) Assuming all donors are ionized, calculate the room temperature current flow if contacts are made on the two ends of the bar and 10 V is applied to the bar.
   (b) Find the electric field in the bar for the conditions of (a).
   (c) What fraction of the current flows in the form of hole current for the conditions of (a)?
   (d) Find the resistivity of the silicon.
   (e) If the silicon were replaced by gallium phosphide and the doping was still $N_d = 5 \times 10^{16}$ cm^{-3} repeat (a), (b) and (c).
   (f) If the silicon temperature was increased to 120°C, repeat (a), (b) and c). Assume that carrier mobility and bandgap are not affected by the increase in temperature.

1.6 Now, instead of being uniformly doped, the bar of Problem 1.5 is doped with a linearly increasing donor doping concentration, such that the left end of the bar (LHS) is doped with a concentration of 1×10^{16} cm^{-3} and the right end of the bar (RHS) is doped with a concentration of 1×10^{17} cm^{-3}.
   (a) Determine the doping level at three points in the bar:
      (i) at 3 cm from the LHS;
      (ii) at the midpoint;
      (iii) at 9 cm from the LHS.

(b) Assuming equilibrium conditions (no applied voltage) find the built-in electric field in the bar at each of positions (i), (ii) and (iii). Hint: Find the gradient in the doping about each point.

(c) Find the electron drift current flowing in the bar at positions (i), (ii) and (iii).

(d) Explain how the bar can be in equilibrium given the existence of these electric fields and drift currents.

(e) Find the Fermi level relative to the top of the valence band for each of positions (i), (ii) and (iii).

(f) Sketch the band diagram as a function of position in equilibrium along the length of the bar showing the location of the Fermi energy.

1.7 A square silicon semiconductor sheet $50 \, \text{cm}^2$ in area and $0.18 \, \text{mm}$ in thickness is uniformly doped with both acceptors ($N_a = 5 \times 10^{16} \, \text{cm}^{-3}$) and donors ($N_d = 2 \times 10^{16} \, \text{cm}^{-3}$).

(a) Assuming all donors and acceptors are ionized, calculate the room temperature current flow if the silicon is contacted by metal contact strips that run the full length of two opposing edges of the sheet and $10 \, \text{V}$ is applied across the contacts.

(b) Repeat (a) but assume that the sheet is $100 \, \text{cm}^2$ in area instead of $50 \, \text{cm}^2$. Does the current change with area? Explain.

(c) Is the recombination time τ important for the measurement of current flow? Explain carefully what role recombination plays in the calculation.

1.8 An undoped silicon semiconductor sheet $50 \, \text{cm}^2$ in area and $0.18 \, \text{mm}$ thick is illuminated over one entire surface and an electron-hole pair generation rate of $10^{21} \, \text{cm}^{-3} \, \text{s}^{-1}$ is achieved uniformly throughout the material.

(a) Determine the separation of the quasi-Fermi levels. The carrier lifetime is $2 \times 10^{-6} \, \text{s}$.

(b) Calculate the room temperature current flow if the silicon is contacted by narrow metal contact strips that run the full length of two opposing edges of the sheet and $10 \, \text{V}$ is applied across the contacts under illumination conditions. The carrier lifetime is $2 \times 10^{-6} \, \text{s}$.

(c) Explain how a higher/lower recombination time would affect the answer to (b). How does the recombination time of $2 \times 10^{-6} \, \text{s}$ compare with the transit time of the carriers, which is the time taken by the carriers to traverse the silicon sheet from one side to the other side? This silicon sheet is functioning as a *photoconductive* device since its conductivity depends on illumination. If the transit time is small compared to the recombination time then *gain* can be obtained since more than one carrier can cross the photoconductive sheet before a recombination event takes place on average. Gains of 100 or 1000 may be obtained in practice in photoconductors. How long a recombination time would be required for a gain of 10 to be achieved?

1.9 A sample of n-type silicon is doped to achieve E_F at $0.3 \, \text{eV}$ below the conduction band edge at room temperature.

(a) Find the doping level.

(b) The n-type silicon sample is in the form of a square cross-section bar at room temperature and it carries a current of $3 \times 10^{-8} \, \text{Å}$ along its length. If the bar is

10 cm long and has a voltage difference of 100 V end-to-end, find the cross-section dimension of the bar.

1.10 A silicon sample is uniformly optically excited such that its quasi-Fermi level for electrons F_n is 0.419 eV above its quasi-Fermi level for holes F_p. The silicon is n-type with donor concentration $N_D = 1 \times 10^{14}$ cm^{-3}. Find the optical generation rate.

1.11 An n-type silicon wafer is 5.0 mm thick and is illuminated uniformly over its surface with blue light, which is absorbed very close to the silicon surface. Assume that a *surface generation rate* of holes of 3×10^{18} cm^{-2} s^{-1} is obtained over the illuminated surface, and that the excess holes are generated at the silicon surface.
   (a) Calculate the hole concentration as a function of depth assuming a hole lifetime of 2×10^{-6} s. Assume that the hole lifetime is independent of depth.
   (b) Calculate the hole diffusion current as a function of depth.
   (c) Calculate the recombination rate of holes as a function of depth.
   (d) Explain the difference between the surface generation rate of holes and the re-combination rate of holes near the surface. These two rates have different units. Your explanation should refer to the distinction between the units employed.

1.12 Find:
   (a) The n-type doping level required to cause silicon at room temperature to have electrical conductivity 100 times higher than intrinsic silicon at room temperature.
   (b) The p-type doping level required to cause silicon at room temperature to have p-type conductivity 100 times higher than intrinsic silicon at room temperature.

1.13 Intrinsic silicon is uniformly illuminated with 10^{14} photons cm^{-2} s^{-1} at its surface. Assume that each photon is absorbed very near the silicon surface, and generates one electron-hole pair.
   (a) Find the flux of electrons at a depth of 3 μm. Make and state any necessary assumptions.
   (b) Find the total excess electron charge stored in the silicon, assuming the silicon sample is very thick.

1.14 If the Fermi energy in an n-type silicon semiconductor at 300°C is 0.08 eV below the conduction band, and the donor level is 0.02 eV below the conduction band, then find the probability of ionization of the donors.

1.15 In a hypothetical semiconductor with effective masses of electrons and holes the same as silicon, and E_g as in silicon, the conductivity of a rod of the material (measuring 1 mm in diameter and 40 mm in length) from end to end is measured as 4×10^{12} ohms at 300 K.
   (a) If the electron mobility is five times larger than the hole mobility in this material then find the electron mobility. Make and state any necessary assumptions.
   (b) Find the current flow due to the holes only in the rod if a voltage of 10 V is applied across the ends of the rod. Make and state any necessary assumptions.

1.16 An intrinsic, planar, room-temperature silicon sample is exposed to a steady flux of light at its surface. The electron concentration as a result of this is measured to be 100 times higher than n_i, the intrinsic equilibrium concentration, at a depth of 100 μm below the silicon surface. You may use the low-level injection approximation.
   (a) Assuming that the light is all absorbed very near the silicon surface, and that every incident photon excites one EHP, find the total photon flux.

(b) Find the quasi-Fermi level for electrons relative to the Fermi level without the flux of light at a depth of 100 μm.

1.17 A famous experiment that involves both the drift and the diffusion of carriers in a semiconductor is known as the Haynes–Shockley experiment. Search for the experimental details of this experiment and answer the following:

   (a) Make a sketch of the semiconductor sample used in the experiment as well as the location and arrangement of electrodes and the required voltages and currents as well as the connections of electrodes for the appropriate measurements to be made.

   (b) Sketch an example of the time dependence of the output of the experiment.

   (c) Explain how the Einstein relation can be verified using these data.

1.18 A silicon sample is uniformly optically excited such that its quasi-Fermi level for electrons F_n is 0.419 eV above its quasi-Fermi level for holes F_p. The silicon is n-type with donor concentration $N_d = 1 \times 10^{17}$ cm^{-3}. Find the optical generation rate.

1.19 A flash of light at time $t = 0$ is uniformly incident on all parts of a p-type silicon sample with doping of 5×10^{17} cm^{-3}. The resulting EHP concentration is 2×10^{16} EHP cm^{-3}. Find the time-dependent electron and hole concentrations for time t greater that zero.

1.20 Carriers are optically generated at an intrinsic silicon surface. The generation rate is 2×10^{19} EHP/cm^2 s. Assume that all the photons are absorbed very close to the silicon surface. Find the diffusion current of electrons just below the surface, and state clearly the assumptions you used to obtain the result.

1.21 A silicon sample is doped with 6×10^{16} donors cm^{-3} and N_a acceptors cm^{-3}. If E_f lies 0.4 eV below E_f in intrinsic silicon at 300 K, find the value of N_a.

1.22 Electric current flows down a silicon rod 1 cm in length and 0.3 mm in diameter. The silicon is n-type with $N_a = 1 \times 10^{17}$ cm^{-3}. A potential difference of 10 V is applied to the rod end-to-end. How many electrons drift through the rod in 60 seconds?

1.23 Find the energy difference between F_n (quasi-Fermi level for electrons) and E_f if illumination were not present, for a silicon sample in equilibrium containing 10^{15} donors cm^{-3}. Assume room temperature. The bar is uniformly optically excited such that 10^{18} EHP/cm^3 s are generated.

1.24 (a) Find the surface recombination velocity of electrons at a p-type silicon semiconductor surface having the following parameters:

$$p_0 = 1 \times 10^{17} \text{ cm}^{-3}$$
$$\delta n = 1 \times 10^{16} \text{ cm}^{-3} \text{ at the surface.}$$

   Use the diffusion constant for silicon in Appendix 2. Make and state any necessary assumptions.

   (b) Sketch a band diagram as a function of distance x from the semiconductor surface to a few diffusion lengths away from the surface.

   (c) Repeat (a) and (b) for a GaAs semiconductor with the same parameters.

1.25 (a) Find the surface recombination velocity of holes at an n-type silicon semiconductor surface having the following parameters:

$$n_0 = 1 \times 10^{16} \text{ cm}^{-3}$$
$$\delta p = 1 \times 10^{17} \text{ cm}^{-3} \text{ at the surface.}$$

Use the diffusion constant for silicon in Appendix 2. Make and state any necessary assumptions.

(b) Sketch a band diagram as a function of distance x from the semiconductor surface to a few diffusion lengths away from the surface.

(c) Repeat (a) and (b) for a GaAs semiconductor with the same parameters.

2

The PN Junction Diode

Objectives

1. Understand the structure of a p-n junction.
2. Obtain a qualitative understanding of diode current flow and the roles of drift and diffusion currents.
3. Derive a quantitative model of diode contact potential based on the band model.
4. Justify and obtain a quantitative model of the depletion region extending away from the junction.
5. Derive a quantitative model of diode current flow based on carrier drift and diffusion and a dynamic equilibrium in the depletion region.

6. Understand phenomena involved in reverse breakdown of a diode and introduce the Zener diode.
7. Understand the tunnel diode based on a band model and its significance in solar cell applications.
8. Understand generation and recombination currents that exist in the depletion region of a p-n junction.
9. Introduce the physics of metal semiconductor contacts including the Schottky diode and ohmic contacts.
10. Introduce the heterojunction, its band structure, and its relevance to optoelectronic devices such as solar cells and LEDs.
11. Understand p-n junction capacitance and minority charge storage effects relevant to AC and transient behaviour.

2.1   Introduction

A semiconductor device comprising a *p-n junction diode* is illustrated in Figure 2.1. There are several basic features of the diode including the following:

(a) a metal *anode* contact applied to a p-type semiconductor forming a metal-semiconductor junction;
(b) a p-type semiconductor;
(c) a p-n semiconductor junction between p-type and n-type semiconductors;
(d) an n-type semiconductor;
(e) a metal *cathode* contact applied to an n-type semiconductor forming another metal-semiconductor junction.

The purpose of this chapter is to cover the principles of the p-n junction diode. Originally the semiconductor diode was used to provide current flow in one direction and current blocking in the other direction. It found widespread application in early logic circuits as a clamping device and as a logic adder. It is widely used as a rectifier in power supplies and as a detector in radio circuits. Two of these circuit applications are illustrated in Figure 2.1. The applications of specific interest to this book are the more recent derivatives of the semiconductor diode that have become widely used semiconductor diode devices for light emission (LEDs) and for solar power generation (solar cells).

The principles underlying diode operation in general will be covered in this chapter.

Solar cells and LEDs can be fabricated from a range of both inorganic and organic semiconductor diode materials. Chapters 4, 5 and 6 will cover device designs and specific fabrication methods, and build on these principles to further develop topics that are specific to solar cells and LEDs based on inorganic as well as organic semiconductors. A good understanding of all diode-based semiconductor devices requires a fundamental understanding of the p-n junction diode including the electrical contacts made to the diode.

We will begin with the *abrupt junction* semiconductor p-n diode in which the transition from n-type material to p-type material occurs abruptly. This is achieved by a step change in the doping species on either side of the semiconductor junction. In practice, it is possible
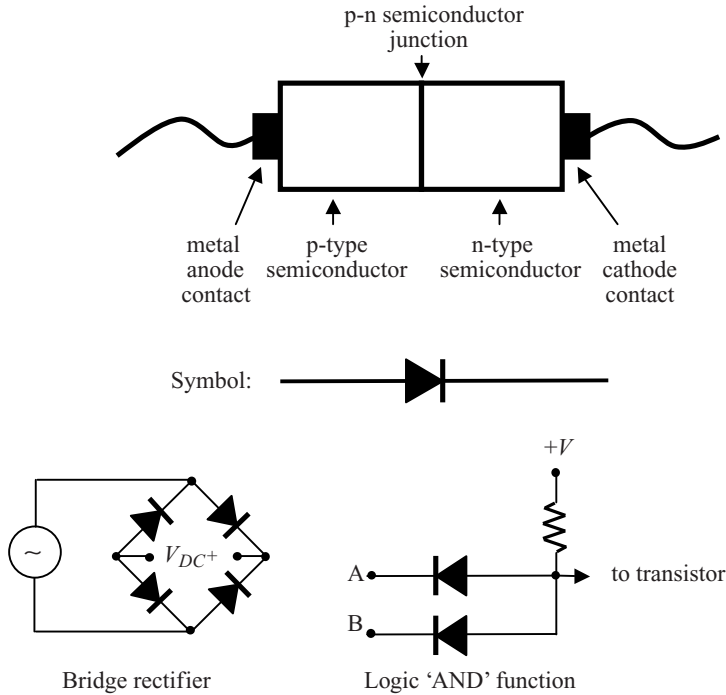
p-n semiconductor junction

metal anode contact   p-type semiconductor   n-type semiconductor   metal cathode contact

Symbol:

V_{DC}^+

Bridge rectifier

$+V$

A
B
to transistor

Logic 'AND' function

Figure 2.1   *The p-n junction diode showing metal anode and cathode contacts connected to semiconductor p-type and n-type regions respectively. There are two metal-semiconductor junctions in addition to the p-n semiconductor junction. The diode symbol and two examples of diode applications in circuit design are shown. The diode logic gate was used in early* diode-transistor logic *solid state computers popular in the 1960s but has been replaced by transistor-based designs that consume less power and switch faster*

to make such a transition over a distance of just one atomic layer of the semiconductor crystal. The abrupt junction diode is the easiest diode to model and understand; however, the concepts can be extended to diodes in which the doping transition is gradual rather than abrupt.

The band model is used to describe electron and hole behaviour for the p-n junction and is shown for equilibrium conditions in Figure 2.2. Since the n-type and p-type semiconductor regions are in equilibrium with each other, the Fermi energy is constant. Note the difference in position for the valence and conduction band energies on the two sides of the junction. There is a *transition region* in which the energy bands are sloped to provide a continuous conduction band and a continuous valence band extending from the p-side to the n-side of the junction. The transition region is present even though the p-n junction is abrupt.

We have seen in Section 1.14 that band slope is evidence for an electric field within the semiconductor, which is referred to as a built-in electric field. The origin of this transition region and the associated electric field will be further described in Section 2.3. The direction
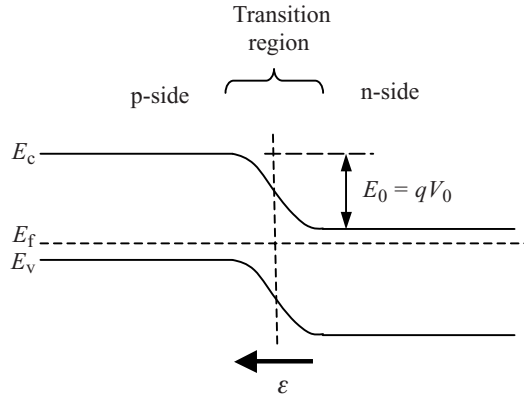
Figure 2.2   *Band model of p-n junction in equilibrium showing constant Fermi energy and transition region to allow valence band and conduction band to be continuous*

of the field is shown in Figure 2.2 to be consistent with the direction of the band slope (see Section 1.14). The built-in field drives electron and hole drift currents across the junction. The change in energy across the p-n junction is labelled E_0 provided equilibrium conditions apply and is referred to as an *energy barrier*. Since the energy gained or lost by a charge may be obtained by multiplying the change in potential by the charge, we can write

$$E_0 = qV_0 \tag{2.1}$$

where V_0 is the *contact potential* of the p-n junction.

2.2   Diode Current

We can now describe the currents that flow in equilibrium. The built-in field ε causes carriers in the transition region to drift. In addition there are substantial carrier concentration gradients across the junction. For electrons, the high electron (majority carrier) concentration in the n-side falls to a low electron (minority carrier) concentration in the p-side. This electron concentration gradient as well as an analogous hole concentration gradient drive diffusion currents across the junction.

There are now four currents to consider:

$I_{n\ drift}$ Electrons (minority carriers) on the p-side that enter the transition region will drift to the right towards the n-side. This current is driven by the built-in electric field.

$I_{p\ drift}$ Holes on the n-side (minority carriers) that enter the transition region will drift to the left towards the p-side. This current is driven by the built-in electric field.

$I_{n\ diffusion}$ Electrons on the n-side (majority carriers) will diffuse to the left. This current is driven by the electron concentration gradient.

$I_{p\ diffusion}$ Holes on the p-side will diffuse to the right. This current is driven by the hole concentration gradient.

p-side   n-side

$I_{\text{n diffusion}}$ $I_{\text{n drift}}$

E_{c}

E_{f}
E_{v}

$I_{\text{p diffusion}}$
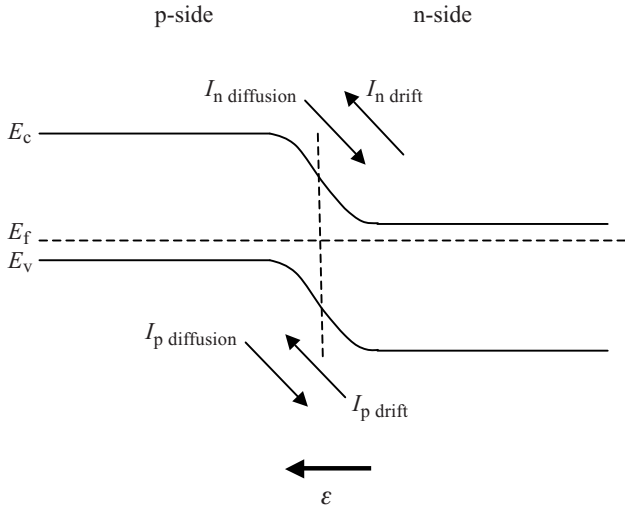
$I_{\text{p drift}}$

ε

Figure 2.3 *Flow directions of the four p-n junction currents. The two diffusion currents are driven by concentration gradients of electrons or holes across the junction and the two drift currents are driven by the electric field. Note that the electron currents flow in the direction opposite to the flux or flow of electrons. The electron diffusion flux is to the left and the electron drift flux is to the right*

These four currents can be viewed schematically in Figure 2.3. The electron currents are reversed in direction from the directions of electron flow since electrons carry negative charge.

If the p-n junction is in equilibrium we can conclude that the following equalities apply:

$$I_{\text{p drift}} + I_{\text{p diffusion}} = 0 \tag{2.2a}$$

$$I_{\text{n drift}} + I_{\text{n diffusion}} = 0 \tag{2.2b}$$

We will show in Section 2.5 that both drift and diffusion currents may be very large; however, the *net* current is the observed diode current.

If we apply a voltage to the diode by connecting an external voltage source to the p-n junction, the currents will no longer cancel out, and the diode is no longer in an equilibrium state. This is illustrated in Figure 2.4. This external voltage is called a *bias* voltage.

Let us first consider the application of a *forward bias* with $V > 0$ in which the p-side is connected to the positive output of the voltage source and the n-side to the negative output. The applied voltage V will fall across the transition region of the p-n junction and will *decrease* the energy barrier height as well as the electric field ε as shown in Figure 2.5. The decrease in barrier height will result in a net current because the opposing drift current will no longer be sufficient to cancel out all the diffusion current. The *net current flow results from a net majority carrier diffusion current* to become

$$I = I_{\text{p diffusion}} + I_{\text{n diffusion}} - I_{\text{p drift}} - I_{\text{n drift}} > 0 \tag{2.3}$$
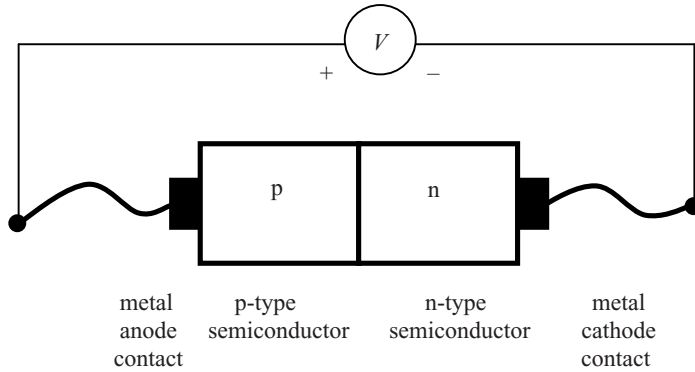
Figure 2.4   *A p-n junction diode with external voltage source connected. The external bias voltage will modify the built-in electric field*

Note that in Figure 2.3 the electron and hole diffusion currents flow in the same direction and may therefore be added in together in Equation 2.3 to obtain the total diode current, whereas hole and electron diffusion *fluxes* flow in opposite directions but they carry opposite charge polarities. The electron and hole drift currents also flow in the same direction and are both negative in Equation 2.3.

   If we now consider the application of a *reverse bias* with $V < 0$ the applied voltage will again fall across the transition region of the p-n junction, which will *increase* the magnitude of both the potential barrier and ε, as shown in Figure 2.6. The increase in the energy barrier will cause drift current to effectively oppose diffusion current. There is,
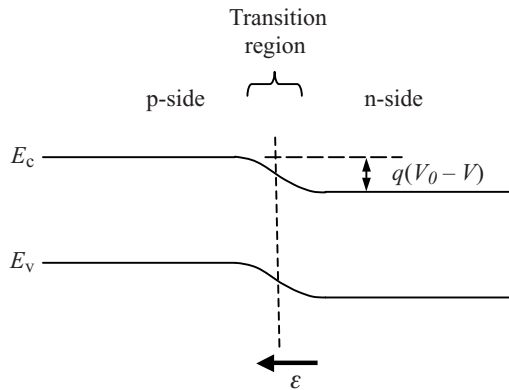
Figure 2.5   *Diode band model with the application of a forward bias. The energy barrier across the transition region is smaller resulting in much higher currents dominated by diffusion currents. In the depletion region ε will be smaller and drift currents no longer compensate for diffusion currents. Note that the applied voltage V (in volts) must be multiplied by the electron charge q (in coulombs) to obtain energy (in joules)*
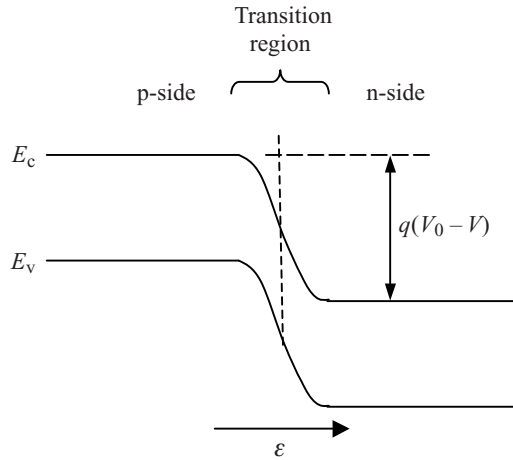
Figure 2.6   *Diode band model with the application of a reverse bias. Since the applied voltage V is negative, the energy barrier as well as electric field ε become larger across the transition region virtually eliminating diffusion currents*

however, a remaining current due to thermally generated *minority* carriers. This constitutes a small net minority carrier drift current and it is assisted by the electric field. The *net current flow is dominated by thermally generated minority carrier drift currents*. From Equation 2.2 we obtain

$$I = I_{\text{p diffusion}} + I_{\text{n diffusion}} - I_{\text{p drift}} - I_{\text{n drift}} < 0$$

The total current I is now small and virtually independent of applied voltage V because I is controlled by the supply of thermally generated minority carriers available to drift and the magnitude of ε is not important. This is analogous to varying the height of a waterfall in a river – the amount of water flowing down the waterfall will depend on the available flow of the water approaching the waterfall and will not be affected by the height of the waterfall. The magnitude of this current is known as the *reverse saturation current*, I_0, and hence I_0 is the net thermally generated drift current supplied by minority carriers.

The diode current may now be plotted as a function of the applied voltage, as shown in Figure 2.7. We will treat diode current quantitatively in Section 2.5.

2.3   Contact Potential

We can calculate the contact potential V_0 using our understanding of energy bands. From Equation 1.35:

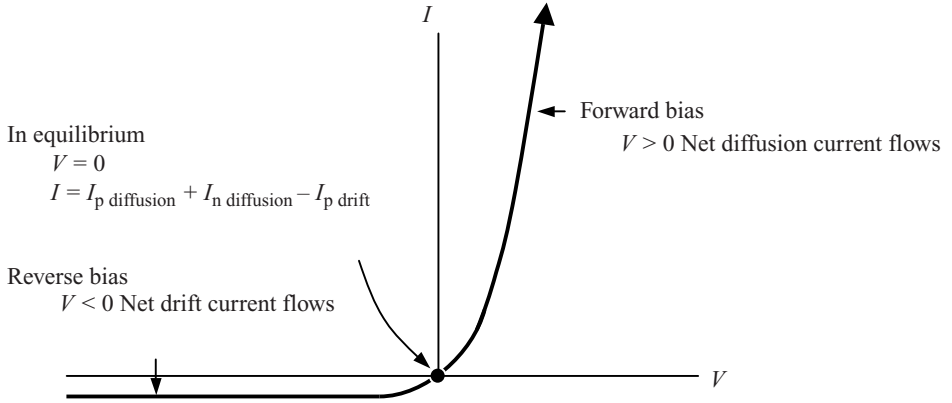$$p_0 = N_v \exp\left(\frac{-(E_f - E_v)}{kT}\right)$$

Figure 2.7   *Diode current as a function of applied voltage. The reverse drift current saturates to a small value and is called the reverse saturation current. When V = 0 the drift and diffusion currents are equal in magnitude and the net current is zero*

We can apply Equation 1.35 to both the n-side and the p-side of the junction:

$$(E_f - E_v)_{p-side} = kT \ln \left(\frac{N_v}{p_0} \right) = kT \ln \left(\frac{N_v}{p_p} \right) \tag{2.4a}$$

where p_p is the equilibrium hole concentration on the p-side. On the n-side,

$$(E_f - E_v)_{n-side} = kT \ln \left(\frac{N_v}{p_0} \right) = kT \ln \left(\frac{N_v}{p_n} \right) \tag{2.4b}$$

where p_n is the equilibrium hole concentration on the n-side. Subtracting Equation 2.4a from Equation 2.4b we obtain

$$E_0 = (E_v)_{p-side} - (E_v)_{n-side} = kT \ln \left(\frac{p_p}{p_n} \right) \tag{2.5}$$

This is illustrated in Figure 2.8.

Now, Equation 2.5 can be expressed in terms of the contact potential V_0 using Equation 2.1, and therefore

$$V_0 = \frac{kT}{q} \ln \left(\frac{p_p}{p_n} \right) \tag{2.6}$$

We can also express this in terms of the doping levels on either side of the junction. From Equations 1.43, 1.44 and 2.6

$$V_0 = \frac{kT}{q} \ln \left(\frac{N_a N_d}{n_i^2} \right) \tag{2.7}$$
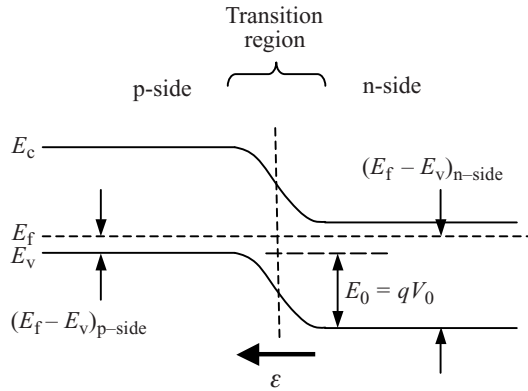
Figure 2.8 *The equilibrium p-n junction energy barrier height E_0 may be obtained from $(E_f - E_v)_{n-side} - (E_f - E_v)_{p-side}$ resulting in Equation 2.5*

The contact potential may also be obtained using electron concentrations. Since $N_a N_d \cong p_p n_n$ and $n_p = \frac{n_i^2}{p_p}$ and $p_n = \frac{n_i^2}{n_n}$ we can also write

$$V_0 = \frac{kT}{q} \ln\left(\frac{n_n}{n_p}\right) = \frac{kT}{q} \ln\left(\frac{p_p}{p_n}\right) \tag{2.8}$$

It is therefore possible to design a p-n junction having a specific built-in potential V_0 by controlling the doping levels in the p-type and n-type regions.

Example 2.1

An abrupt silicon p-n junction diode is doped with $N_a = 1 \times 10^{17}$ cm^{-3} on the p-side and $N_d = 1 \times 10^{17}$ cm^{-3} on the n-side.

Find the built-in potential and sketch the band diagram in equilibrium at room temperature. Include the Fermi level.

Solution

$$V_0 = \frac{kT}{q} \ln\left(\frac{N_a N_d}{n_i^2}\right) = (0.026 \text{ V}) \ln\left[\frac{10^{17} \text{ cm}^{-3} \times 10^{17} \text{ cm}^{-3}}{\left(1.5 \times 10^{10} \text{ cm}^{-3}\right)^2}\right] = 0.817 \text{ V}$$

On the p-side from Example 1.2,

$$N_v = 1.06 \times 10^{19} \text{ cm}^{-3}$$

Now,

$$(E_f - E_v)_{p-side} = kT \ln \left(\frac{N_v}{p_p} \right)$$

$$= (0.026 \text{ eV}) \ln \left(\frac{1.06 \times 10^{19} \text{ cm}^{-3}}{10^{17} \text{ cm}^{-3}} \right) = 0.12 \text{ eV}$$

p-side   n-side

E_c

$E_g = 1.11$ eV

E_f

E_v

$E_0 = qV_0 = 0.817$ eV

$(E_f - E_v)_{p-side} = 0.12$ eV

2.4   The Depletion Approximation

A detailed view of the transition region of Figure 2.2 is redrawn in Figure 2.9. In the central part of the transition region the Fermi energy is close to the middle of the energy gap, which implies that the semiconductor will behave much like intrinsic material even though it is doped either p-type or n-type. This is shown as a *strongly depleted* region in Figure 2.9 because the concentration of charge carriers here is very low as it would be in an intrinsic semiconductor. As we move to the left or right of this region the carrier concentrations gradually return to their normal p-type or n-type equilibrium values respectively, and the energy bands become horizontal lines.

Transition
region

p-side   Strongly   n-side
depleted

E_c

E_f
E_v

Figure 2.9   *Depletion occurs near the junction. In order to establish equilibrium conditions, electrons and holes recombine and the Fermi energy lies close to the middle of the bandgap in the strongly depleted region*

To understand this depletion it is necessary to consider two separate semiconductors, one n-type and one p-type. If they are brought together, electrons in the n-type material and holes in the p-type material close to the junction will quickly diffuse across the junction and annihilate each other by recombination leaving a deficit of holes and electrons. This diffusion and recombination will be complete after a very short time, and then equilibrium conditions will be established. Electrons and holes further away from the junction will also have a chance to diffuse across the junction; however, the electric field that is built up near the junction opposes this and also causes drift currents and the resulting band diagram now is correctly described by Figure 2.2. The equilibrium currents that continue to flow are described by Equation 2.2.

A simplification called the *depletion approximation* is used to model the depletion of charges. We assume complete depletion of charge carriers in a *depletion region* of width W_0 at the junction, and then assume that the carrier concentrations abruptly return to their equilibrium levels on either side of the depletion region as shown in Figure 2.10. Carrier concentrations do not make abrupt concentration changes in real materials or devices and instead gradients in carrier concentrations exist; however, the depletion approximation dramatically simplifies the quantitative model for the p-n junction and the results are highly relevant to real diodes. In Section 2.5 we will consider a more detailed view of these carrier gradients.

Normally a doped semiconductor consists of both mobile carriers and non-mobile ionized dopants as discussed in Section 1.13. The net charge density in the semiconductor is zero since the dopant ions have a charge density that is equal in magnitude and opposite in sign to the mobile charges that they provide.
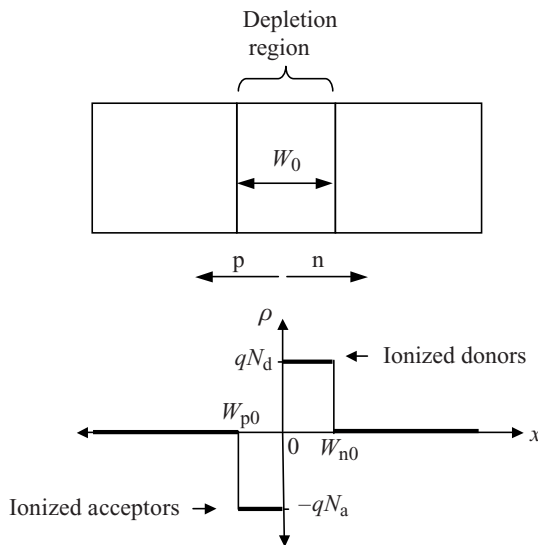
Figure 2.10   *A depletion region of width W_0 is assumed at the junction. Charge density ρ is zero outside of the depletion region. Inside the depletion region a net charge density due to ionized dopants is established. The origin of the x-axis is placed at the junction for convenience*

The semiconductor material within the depletion region of a p-n junction is depleted of carriers but *it is still doped either p-type or n-type* and dopant ions are therefore present in this region. This means that there is a net fixed (non-mobile) charge density in the depletion region from the ionized dopants.

In the p-side the ions are negative. For example, in silicon with p-type aluminium doping the aluminium ions are negatively charged having accepted an electron to yield a valence band hole. In the n-side the ions are positive. Conversely, in silicon with n-type phosphorus doping the phosphorus ions would be positively charged having donated an electron to the conduction band. Once these holes and electrons recombine the depletion region will be left with Al^- ions and P^+ ions having concentrations of N_a and N_d respectively.

Charge densities $-qN_a$ and $+qN_d$ (coulombs per cm^3) will reside on the p-side and n-side of the depletion region respectively, as indicated in Figure 2.10. The magnitude of depletion charge on either side of the junction must be the same since one negative and one positive ion results from each recombination of a hole and an electron. Hence the magnitude of the charge present on either side of the junction is

$$Q = qN_aW_{p0} A = qN_dW_{n0} A \tag{2.9}$$

where W_{p0} and W_{n0} are the widths of the depletion regions on the p- and n-sides of the junction in equilibrium respectively, and A is the cross-sectional area of the diode. From Equation 2.9 we can write

$$\frac{W_{p0}}{W_{n0}} = \frac{N_d}{N_a} \tag{2.10}$$

The charged regions in the depletion region give rise to an electric field. This is the same field that causes the energy bands to tilt in Figure 2.2, and we can now calculate it. Using Gauss's law we can enclose the negative charge on the p-side of the depletion region with a Gaussian surface having surface area A, as shown in Figure 2.11.
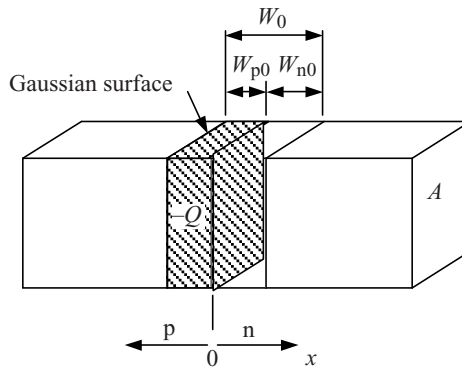
Figure 2.11   *A Gaussian surface having volume AW_{p0} (shaded) encloses the negative charge of magnitude Q on the p-side of the depletion region*

Gauss's law relates the surface integral of electric field ε for a closed surface to the enclosed charge Q or

$$\oint_S \vec{\varepsilon} \cdot \vec{\mathrm{d}s} = \frac{Q}{\epsilon} \tag{2.11}$$

If we assume our Gaussian volume has a relatively large cross-sectional surface area A, and a small depth W_{p0}, then we can approximate the total surface area as $2A$. Since there is symmetry and the two cross-sectional areas A are equivalent, we can use Equation 2.11 to write

$$2\varepsilon_p A = \frac{Q}{\epsilon_0 \epsilon_r}$$

where ε_p is the magnitude of the electric field caused by the depletion charge on the p-side and ε_r is the relative permittivity of the semiconductor, and therefore

$$\varepsilon_p = \frac{Q}{2\epsilon_0 \epsilon_r A} \tag{2.11a}$$

The same reasoning may be applied to a Gaussian surface that encloses depletion charge on the n-side. This will give rise to a field of magnitude

$$\varepsilon_n = \frac{Q}{2\epsilon_0 \epsilon_r A} \tag{2.11b}$$

Since electric field is a vector quantity the relevant electric field directions are shown in Figure 2.12, and it is clear that the total field in equilibrium at the junction is the vector sum of $\overrightarrow{\varepsilon_p}$ and $\overrightarrow{\varepsilon_n}$ yielding the *equilibrium electric field at the semiconductor junction*

$$\varepsilon_0 = -\frac{Q}{\epsilon_0 \epsilon_r A}$$

which may be rewritten using Equation 2.9 as

$$\varepsilon_0 = -\frac{q N_d W_{n0}}{\epsilon_0 \epsilon_r} = -\frac{q N_a W_{p0}}{\epsilon_0 \epsilon_r} \tag{2.12}$$
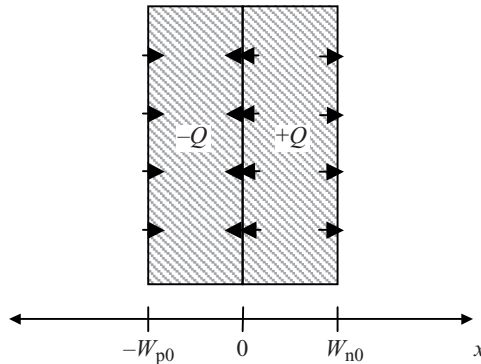
Figure 2.12 *The electric field directions for the two parts of the depletion region showing that the fields add at the junction*

The minus sign indicates that the field points in the negative x direction. The field at other points along the x-axis may also be evaluated. We have assumed that the cross-sectional area of the junction is very large compared to the depletion width. Since the electric field due to an infinite plane of charge is independent of the distance from the plane, the vector quantities in Figure 2.12 will cancel out and the net field will be zero at $x = -W_{p0}$ and at $x = -W_{n0}$. At other x values, the electric field will vary linearly as a function of x and may be calculated by appropriately applying Gauss's law (see Problem 2.5). The resulting electric field will also give rise to a potential

$$V_0(x) = -\int_{-W_{p0}}^{W_{n0}} \varepsilon(x)\mathrm{d}x \tag{2.13}$$

Both $\varepsilon(x)$ and $V(x)$ are shown in Figure 2.13, which also shows that a contact potential V_0 results directly from the depletion model. This contact potential V_0 is the same quantity that we introduced in Section 2.3.
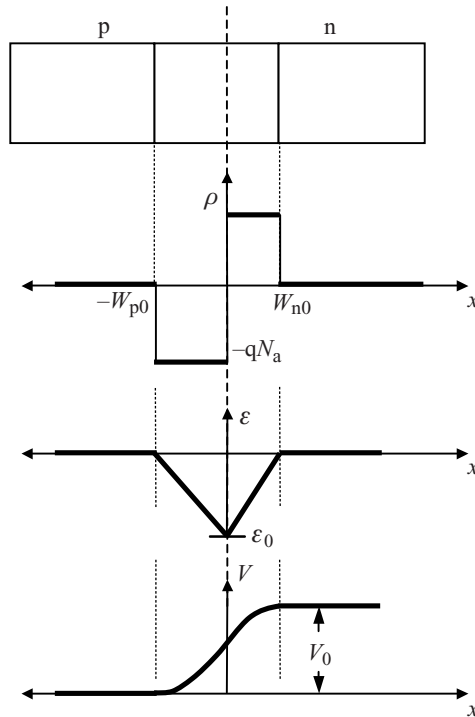
Figure 2.13   *The equilibrium electric field $\varepsilon(x)$ and potential $V(x)$ for the p-n junction follow from the application of Gauss's law to the fixed depletion charge. Note that V on the n-side is higher compared to the p-side, whereas in Figure 2.2 the energy levels on the n-side are lower. This is the case because the energy scale in Figure 2.2 is for* electron *energy levels; however, the voltage scale in Figure 2.13 is established for a* positive *charge by convention*

The integral (Equation 2.13) is the area under the ε versus x curve in Figure 2.13, which may be calculated using the area of a triangle, and we obtain

$$V_0 = \frac{1}{2}bh = \frac{1}{2}W_0\varepsilon_0$$

and using Equation 2.12 to express ε_0,

$$V_0 = \frac{qN_dW_{n0}}{2\epsilon_0\epsilon_r}W_0 = \frac{qN_aW_{p0}}{2\epsilon_0\epsilon_r}W_0$$

Since $W_0 = W_{n0} + W_{p0}$,

$$V_0 = \frac{q}{2\epsilon_0\epsilon_r}\frac{N_aN_d}{N_a + N_d}W_0^2 \tag{2.14}$$

or

$$W_0 = \sqrt{\frac{2\epsilon_0\epsilon_r V_0}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)} \tag{2.15a}$$

with

$$W_{p0} = \frac{W_0N_d}{N_a + N_d} \tag{2.15b}$$

and

$$W_{n0} = \frac{W_0N_a}{N_a + N_d} \tag{2.15c}$$

It is interesting to note that the depletion approximation is consistent with an externally applied voltage falling across the depletion region. Since the depletion region has high resistivity, and the neutral regions on either side of it have high conductivities, we are justified in stating that applied reverse-bias voltages will drop across the depletion region.

Example 2.2

(a) Find the depletion layer width in both the n-side and the p-side of the abrupt silicon p-n junction diode of Example 2.1 doped with $N_a = 10^{17}$ cm^{-3} on the p-side and $N_d = 10^{17}$ cm^{-3} on the n-side. Find the equilibrium electric field at the semiconductor junction ε_0. Sketch the electric field and the potential as a function of position across the junction.

(b) Repeat if the doping in the p-side is increased to $N_a = 1 \times 10^{18}$ cm^{-3} and the doping in the n-side is decreased to $N_d = 1 \times 10^{16}$ cm^{-3}. This is called a p$^+$-n junction to indicate the heavy doping on the p-side.

Solution

(a)

$$W_0 = \sqrt{\frac{2\epsilon_0\epsilon_r V_0}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)}$$

$$= \sqrt{\frac{2\left(8.85 \times 10^{-14} \times 11.8 \text{ F cm}^{-1}\right)(0.817 \text{ V})}{1.6 \times 10^{-19} \text{ C}}\left(\frac{1}{10^{17} \text{ cm}^{-3}} + \frac{1}{10^{17} \text{ cm}^{-3}}\right)}$$

$$= 1.51 \times 10^{-5} \text{ cm} = 0.15 \text{ μm}$$

On p-side:

$$W_{p0} = \frac{W_0 N_d}{N_a + N_d} = \frac{0.15 \text{ μm} \times 10^{17} \text{ cm}^{-3}}{10^{17} \text{ cm}^{-3} + 10^{17} \text{ cm}^{-3}} = 0.075 \text{ μm}$$

On n-side:

$$W_{n0} = \frac{W_0 N_a}{N_a + N_d} = \frac{0.15 \text{ μm} \times 10^{17} \text{ cm}^{-3}}{10^{17} \text{ cm}^{-3} + 10^{17} \text{ cm}^{-3}} = 0.075 \text{ μm}$$

$$\varepsilon_0 = -\frac{q N_d W_{n0}}{\epsilon_0 \epsilon_r} = \frac{1.6 \times 10^{-19} \text{ C} \times 10^{17} \text{ cm}^{-3} \times 0.75 \times 10^{-5} \text{ cm}}{8.85 \times 10^{-14} \text{ F cm}^{-1} \times 11.8}$$

$$= 1.15 \times 10^5 \text{ V cm}^{-1}$$

(b)
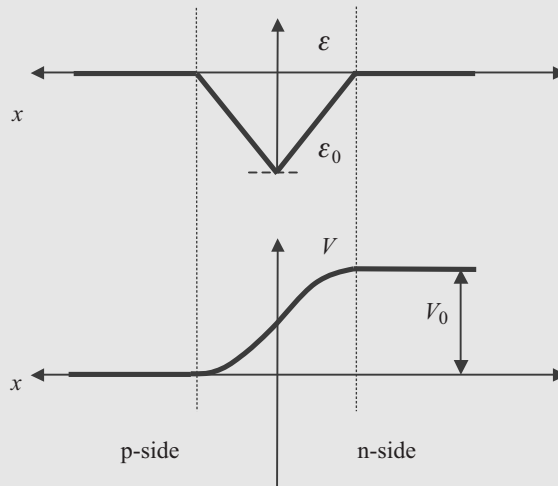
$$V_0 = kT/q \ln\left((N_a N_d)/\left(n_i^2\right)\right)$$

$$= (0.026 \text{ V}) \ln\left[\left(10^{18} \text{ cm}^{-3} \times 10^{16} \text{ cm}^{-3}\right)/\left(1.5 \times 10^{10} \text{ cm}^{-3}\right)^2\right] = 0.817 \text{ V}$$

$$W_0 = \sqrt{\frac{2\epsilon_0\epsilon_r V_0}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)}$$

$$= \sqrt{\frac{2\left(8.85 \times 10^{-14} \times 11.8 \text{ F cm}^{-1}\right)(0.817 \text{ V})}{1.6 \times 10^{-19} \text{ C}}\left(\frac{1}{10^{18} \text{ cm}^{-3}} + \frac{1}{10^{16} \text{ cm}^{-3}}\right)}$$

$$= 3.4 \times 10^{-5} \text{ cm} = 0.34 \text{ μm}$$

On p-side:

$$W_{p0} = \frac{W_0 N_d}{N_a + N_d} = \frac{0.34 \text{ μm} \times 10^{16} \text{ cm}^{-3}}{10^{18} \text{ cm}^{-3} + 10^{16} \text{ cm}^{-3}} \cong 0.33 \times 10^{-2} \text{ μm} = 3.3 \text{ nm}$$

On n-side:

$$W_{n0} = \frac{W_0 N_a}{N_a + N_d} = \frac{0.34 \times 10^{18} \text{ cm}^{-3}}{10^{18} \text{ cm}^{-3} + 10^{16} \text{ cm}^{-3}} \cong 0.33 \text{ μm}$$

$$\varepsilon_0 = -\frac{q N_d W_{n0}}{\epsilon_0 \epsilon_r} = \frac{1.6 \times 10^{-19} \text{ C} \times 10^{16} \text{ cm}^{-3} \times 3.3 \times 10^{-5} \text{ cm}}{8.85 \times 10^{-14} \text{ F cm}^{-1} \times 11.8}$$

$$= 5.1 \times 10^4 \text{ V cm}^{-1}$$

Note that the depletion region is almost entirely in the n-side of the p$^+$-n junction. Almost all the built-in potential V_0 appears in the n-side. The reverse is true for an n$^+$-p junction.
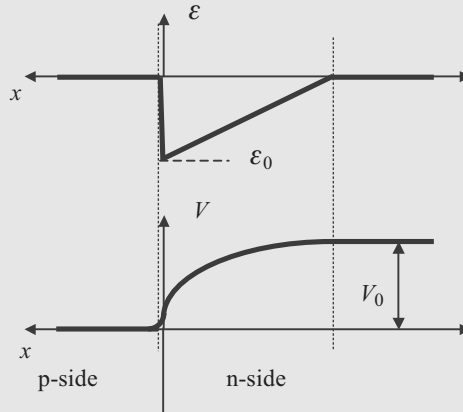
2.5   The Diode Equation

We shall now derive the current–voltage relationship for a p-n junction diode with an externally applied voltage bias, which will result in the *diode equation*. Under equilibrium conditions we saw that diffusion and drift currents cancel out; however, if an external bias is added, then a net current will flow.

In Figures 2.5 and 2.6 we saw that adding an external voltage V creates a new junction potential $V_0 - V$. Since the p-n junction is no longer in equilibrium we cannot continue to use the Fermi energy to describe and derive junction physics; however, we will now show that inside the depletion region of a biased diode, conditions consistent with a *dynamic equilibrium* are maintained even though current flows.

There is considerable confusion about the analysis of diode current in textbooks on semiconductor devices; however, the following picture is essential to a proper understanding of what controls diode current. This picture also allows us to derive the diode equation.

In equilibrium conditions, there are both drift and diffusion currents flowing across the depletion region. These currents cancel out and no net current flows. We will now illustrate, using Example 2.3, that when a diode is biased, very large diffusion and drift currents can flow, and that the range of typical diode currents implies that drift and diffusion currents almost exactly compensate for each other. It is important to remember that the depletion model of a diode is not sufficient to analyse diode current flow since the depletion model appears to predict that no currents flow in a diode. We must therefore remember that gradients in carrier concentrations actually do exist within what we have described as the depletion region.

Example 2.3

(a) For the silicon diode of Example 2.1 assume a junction area of 1 mm^2. Estimate the expected equilibrium majority carrier *diffusion* current across the depletion region if drift current is neglected.
(b) For the silicon diode of Example 2.1 assume a junction area of 1 mm^2. Estimate a typical minority carrier *drift* current if diffusion current is neglected. Assume minority carrier concentrations of only 1% of the majority carrier concentrations.
(c) Explain why these current are not measured externally.

Solution

(a) From Appendix 2, $\mu_n = 1350 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and $\mu_p = 480 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ for silicon. Therefore, using the Einstein relation,

$$D_n = \frac{kT}{q} \mu_n = 0.026 \text{ V} \times 1350 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} = 3.51 \times 10^1 \text{ cm}^2 \text{ s}^{-1}$$

and

$$D_p = \frac{kT}{q} \mu_p = 0.026 \text{ V} \times 480 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} = 1.25 \times 10^1 \text{ cm}^2 \text{ s}^{-1}$$

The depletion model allows us to determine the width of the depletion region. From Example 2.1a, we have majority carrier concentrations on both sides of the depletion region of $1 \times 10^{17} \text{ cm}^{-3}$ and a depletion region of width $W = 0.15 \text{ μm}$. The depletion model predicts physically impossible abrupt changes in majority carrier concentration at the edges of the depletion region. To avoid this difficulty

we will assume that the concentration gradient extends all the way across the depletion region. We also restrict our analysis to diffusion currents and assume there is no electric field in the depletion region, which means that drift currents do not flow. The majority diffusion current would therefore flow due to the large concentration gradient of 1×10^{17} cm^{-3} across the depletion region of width 0.15 μm.

Now,

$$J_n(x)_{\text{diffusion}} = q D_n \frac{dn(x)}{dx}$$

$$= 1.6 \times 10^{-19} \text{ C} \times 3.51 \times 10^1 \text{ cm}^2 \text{ s}^{-1} \frac{1 \times 10^{17} \text{ cm}^{-3}}{1.5 \times 10^{-5} \text{ cm}}$$

$$= 3.74 \times 10^4 \text{ A cm}^{-2}$$

and

$$J_p(x)_{\text{diffusion}} = -q D_p \frac{dp(x)}{dx}$$

$$= 1.6 \times 10^{-19} \text{ C} \times 1.25 \times 10^1 \text{ cm}^2 \text{ s}^{-1} \frac{1 \times 10^{17} \text{ cm}^{-3}}{1.5 \times 10^{-5} \text{ cm}}$$

$$= 1.33 \times 10^4 \text{ A cm}^{-2}$$

For a 1 mm^2 junction area total diffusion current is $374 + 133 = 507$ A. Typical operating diode current densities are *orders of magnitude* smaller than this. A real diode with a junction area of 1 mm^2 would typically carry currents in the range of microamps or milliamps, with a maximum current on the order of only a few amps.

(b) Assume a minority carrier concentration in the depletion region of only 1% of the majority carrier concentration in the diode. Think of these minority carriers as being the result of carrier diffusion. This yields a minority carrier concentration of 1×10^{15} cm^{-3}. We will now neglect diffusion currents and only consider minority carrier drift currents and an *average* electric field inside the depletion region of $\varepsilon = \frac{V_0}{W_0}$.

Now,

$$J_n(x)_{\text{drift}} = q \mu_n n \varepsilon$$

$$= 1.6 \times 10^{-19} \text{ C} \times 1350 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \times 1 \times 10^{15} \text{ cm}^{-3}$$

$$\times \frac{0.817 \text{ V}}{1.5 \times 10^{-5} \text{ cm}} = 1.18 \times 10^4 \text{ A cm}^{-2}$$

and

$$J_p(x)_{\text{drift}} = q \mu_p p \varepsilon$$

$$= 1.6 \times 10^{-19} \text{ C} \times 480 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1} \times 1 \times 10^{15} \text{ cm}^{-3}$$

$$\times \frac{0.817 \text{ V}}{1.5 \times 10^{-5} \text{ cm}} = 4.2 \times 10^3 \text{ A cm}^{-2}$$

> For a 1 mm² junction area total drift current is $118 + 42 = 160$ A. As is part (a), typical operating diode current densities are *orders of magnitude* smaller than this.
>
> (c) Measured diode current is the result of a delicate balance between diffusion and drift current. It is clear from parts (a) and (b) that the depletion region in a diode supports both drift and diffusion currents flowing in opposite directions that almost exactly compensate for each other when typical current densities in semiconductor diode devices flow. This is somewhat analogous to someone running up a high-speed escalator that is heading down. If the runner runs at almost the same speed as the escalator the net speed of the runner is very slow and is much slower than the escalator speed.

Due to the observation from Example 2.3 we are justified in considering that a dynamic equilibrium in the depletion region is a good approximation *even with the application of a bias voltage V* and therefore, for holes, from Equation 1.56,

$$q\mu_p p(x)\varepsilon(x) - qD_p \frac{dp(x)}{dx} \cong 0$$

Solving for $\varepsilon(x)$ we obtain

$$\varepsilon(x) = \frac{D_p}{\mu_p} \frac{1}{p(x)} \frac{dp(x)}{dx}$$

Using the Einstein relation (Equation 1.58) we obtain

$$\varepsilon(x) = \frac{kT}{q} \frac{1}{p(x)} \frac{dp(x)}{dx}$$

Now integrating this across the depletion region and including an applied bias voltage V,

$$V_0 - V = \int_{-W_p}^{W_n} \varepsilon(x)dx = \frac{kT}{q} \int_{-W_p}^{W_n} \frac{1}{p(x)}dp(x) = \frac{kT}{q} \ln\left(\frac{p(-W_p)}{p(W_n)}\right) \qquad (2.16)$$

where $p(-W_p)$ is the majority hole concentration on the p-side of the depletion region and $p(W_n)$ is the minority hole concentration on the n-side of the depletion region.

Substituting for V_0 in Equation 2.16 using Equation 2.6 we obtain

$$\frac{kT}{q} \ln\left(\frac{p_p}{p_n}\right) - V = \frac{kT}{q} \ln\left(\frac{p(-W_p)}{p(W_n)}\right)$$

and rearranging we have

$$-V = \frac{kT}{q} \ln\left(\frac{p(-W_p)}{p(W_n)} \frac{p_n}{p_p}\right)$$

This can be simplified in the case of *low level injection* in which any changes in carrier concentration are small compared to the majority carrier concentrations. This means that $p\left(-W_{p}\right)$ is almost the same as p_{p} and hence

$$-V = \frac{kT}{q} \ln\left(\frac{p_{n}}{p\left(W_{n}\right)}\right)$$

Solving for $p\left(W_{n}\right)$, which is the new minority hole concentration at the edge of the depletion region on the n-side, we obtain

$$p(W_{n}) = p_{n} \exp\left(\frac{qV}{kT}\right) \tag{2.17}$$

Note that if $V = 0$ then $p\left(W_{n}\right) = p_{n}$ as expected. The addition of a bias voltage V therefore multiplies the minority hole concentration at the edge of the depletion region on the n-side by the exponential term $\exp\left(\frac{qV}{kT}\right)$.

The same procedure may be applied to electrons. The integral for electrons from Equation 1.56 is

$$q\mu_{n}n(x)\varepsilon(x) + qD_{n}\frac{dn(x)}{dx} \cong 0$$

From this, using the Einstein relation, we obtain

$$\varepsilon(x) = \frac{-D_{n}}{\mu_{n}}\frac{1}{n(x)}\frac{dn(x)}{dx} = \frac{-kT}{q}\frac{1}{n(x)}\frac{dn(x)}{dx}$$

Now integrating across the depletion region,

$$V_{0} - V = \int_{-W_{p}}^{W_{n}}\varepsilon(x)dx = \frac{-kT}{q}\int_{-W_{p}}^{W_{n}}\frac{1}{n(x)}dn(x) = \frac{-kT}{q}\ln\left(\frac{n(-W_{p})}{n(W_{n})}\right) \tag{2.18}$$

Substituting Equation 2.8 into Equation 2.18 we obtain

$$\frac{kT}{q}\ln\left(\frac{n_{n}}{n_{p}}\right) - V = \frac{-kT}{q}\ln\left(\frac{n(-W_{p})}{n(W_{n})}\right)$$

and therefore

$$V = \frac{kT}{q}\ln\left(\frac{n(-W_{p})}{n(W_{n})}\frac{n_{n}}{n_{p}}\right)$$

Now for low-level injection, the majority electron concentration at the edge of the depletion region on the n-side $n(W_{n})$ is taken to be the same as n_{n}. Hence,

$$V = \frac{kT}{q}\ln\left(\frac{n(-W_{p})}{n_{p}}\right)$$

Solving for $n(-W_p)$ we obtain

$$n(-W_p) = n_p \exp\left(\frac{qV}{kT}\right) \tag{2.19}$$

The minority electron concentration at the edge of the depletion region on the p-side is now exponentially dependent on the applied voltage V, and if $V = 0$ then $n(-W_p) = n_p$ as expected.

   We have now calculated the minority carrier concentrations on either side of the depletion region using the concept of a dynamic equilibrium in the depletion region; however, we have not yet calculated the diode current. We know that the diode current is ultimately limited and controlled by a process other than the dynamic equilibrium. The limiting process in question is the *recombination of the minority carriers* on either side of the depletion region.

   Majority carriers are injected across the depletion region and appear on the other side of the junction where they become minority carriers. These minority carriers then diffuse and eventually recombine with majority carriers in order for current to flow through the diode. Recombination is usually a relatively slow process in semiconductors and even more so in indirect gap semiconductors such as silicon, and it is therefore not surprising that this is a rate-limiting step. Recombination processes are further discussed in Chapters 3 and 5.

   In order to analyse the net current two new coordinates are introduced. The distance to any point in the p-type semiconductor from the depletion region edge will be x_p and the corresponding distance on the n-side will be x_n as shown in Figure 2.14.

   It is useful to express Equations 2.17 and 2.19 in terms of *changes* in minority carrier concentrations, and we will use the new coordinate system. From Equation 2.17,

$$\Delta p(W_n) = \Delta p_{n(x_n=0)} = p_n \exp\left(\frac{qV}{kT}\right) - p_n = p_n\left(\exp\left(\frac{qV}{kT}\right) - 1\right) \tag{2.20a}$$

and similarly from Equation 2.19,

$$\Delta n(-W_p) = \Delta n_{p(x_p=0)} = n_p \exp\left(\frac{qV}{kT}\right) = n_p\left(\exp\left(\frac{qV}{kT}\right) - 1\right) \tag{2.20b}$$

   From Equations 2.20a and 2.20b for forward bias with positive values of V the minority carrier concentrations increase exponentially with V. In reverse bias for large negative values of V the changes in carrier concentrations Δn_p and Δp_n can be negative and can
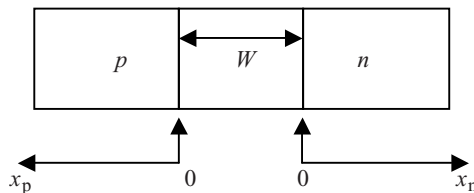
Figure 2.14   *Coordinates x_p and x_n define distances into the p-type and n-type semiconductor regions starting from the depletion region edges*

approach n_p and p_n. This implies that the carrier concentrations of minority carriers decrease to virtually zero at the depletion region edges for reverse bias conditions.

The changes in carrier concentration in Equation 2.20 exist on either side of the depletion region. Far away from the depletion region carrier concentrations will return to their equilibrium levels. This occurs because carriers will be unaffected by the junction if they are several diffusion lengths away (see Figure 1.28), which implies that there are gradients in minority carrier concentration on either side of the depletion region. These concentration gradients will give rise to diffusion currents. The excess minority carrier concentrations exponentially decay with distances x_n and x_p into the neutral regions of the diode.

Another way to view this is using quasi-Fermi levels, which are valid even if there is an applied bias. On the p-side, applying Equation 1.60 and using the low-level injection assumption, we can write

$$F_n - F_p = kT \ln \left(\frac{n_{p(biased)}}{n_p} \right)$$

or

$$\Delta n_{p(x_p=0)} = n_{p(biased)} - n_p = n_p \left(\exp \left(\frac{F_n - F_p}{kT} \right) - 1 \right)$$

Comparing this to Equation 2.20b we can see that the separation between the quasi-Fermi levels is equal to the applied voltage V multiplied by q. The same applies to the n-side, and using Equations 1.60 and 2.20a

$$\Delta p_{n(x_n=0)} = p_{n(biased)} - p_n = p_n \left(\exp \left(\frac{F_n - F_p}{kT} \right) - 1 \right)$$

The resulting quasi-Fermi levels are shown in Figure 2.15.

The diffusion equation for electrons (Equation 1.64b), which describes both diffusion and recombination written in terms of diffusion length $L_n = \sqrt{D_n \tau_n}$ is

$$\frac{d^2 \delta n(x)}{dx^2} = \frac{\delta n(x)}{L_n^2}$$

and may be solved using Equation 2.20 as a boundary condition at $x_p = 0$. For the p-side of the junction the excess electron concentration as a function of x_p becomes

$$\delta n \left(x_p \right) = \Delta n_{p(x_p=0)} \exp \left(-\frac{x_p}{L_n} \right) = n_p \left(\exp \left(\frac{qV}{kT} \right) - 1 \right) \exp \left(-\frac{x_p}{L_n} \right) \quad (2.21a)$$

and similarly on the n-side, the excess hole concentration as a function of x_n is

$$\delta p \left(x_n \right) = \Delta p_{n(x_n=0)} \exp \left(-\frac{x_n}{L_p} \right) = p_n \left(\exp \left(\frac{qV}{kT} \right) - 1 \right) \exp \left(-\frac{x_n}{L_p} \right) \quad (2.21b)$$

Using Equation 1.54 the electron diffusion current in the p-side will be

$$I_n \left(x_p \right) = q A D_n \frac{d \delta n \left(x_p \right)}{dx_p} = -\frac{q A D_n}{L_n} n_p \left(\exp \left(\frac{qV}{kT} \right) - 1 \right) \exp \left(-\frac{x_p}{L_n} \right) \quad (2.21c)$$
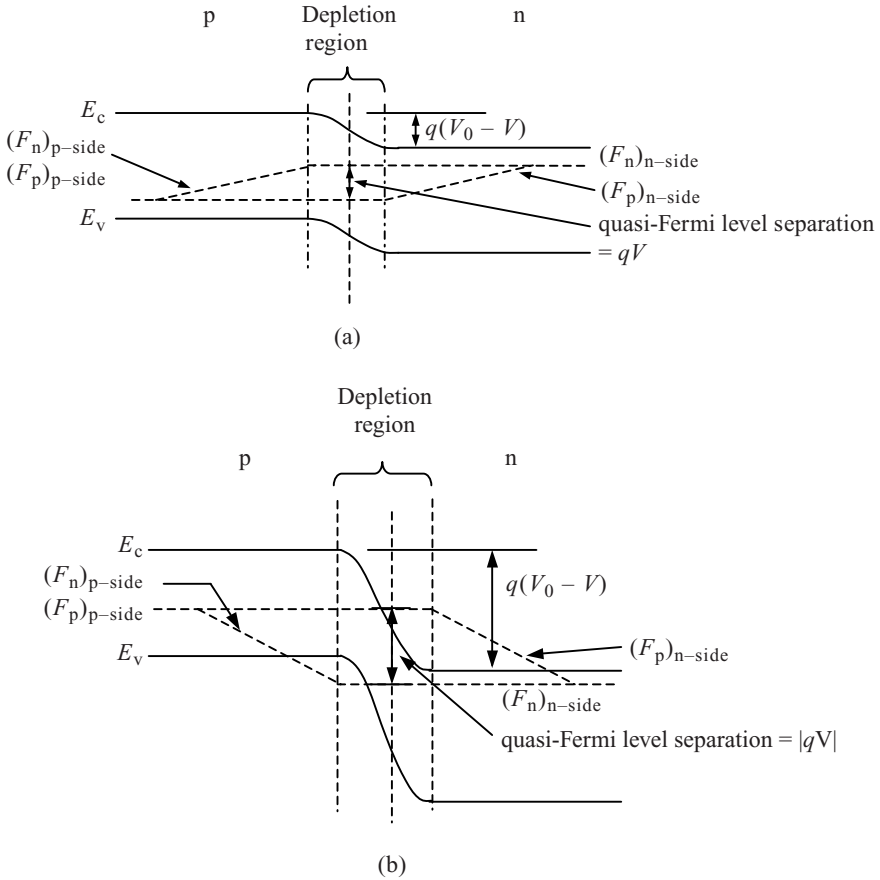
Figure 2.15 *(a) Quasi-Fermi levels for a forward biased junction. $(F_p)_{p-side}$ and $(F_n)_{n-side}$ are horizontal because for low-level injection the majority carrier concentrations are approximately fixed; however, minority carrier concentrations increase towards the depletion region due to carrier injection and therefore $(F_n)_{p-side}$ and $(F_p)_{n-side}$ are tilted. The separation between quasi-Fermi levels at the depletion region edge is equal to qV. (b) Quasi-Fermi levels for a reverse biased junction with V < 0. Minority carriers drift across the depletion region. Minority carrier concentrations now decrease towards the depletion region. Therefore $(F_n)_{p-side}$ and $(F_p)_{n-side}$ are tilted. The separation between quasi-Fermi levels at the depletion region edges is equal to $|qV|$*

and in the n-side,

$$I_p(x_n) = -qAD_p \frac{\mathrm{d}\delta n\,(x_n)}{\mathrm{d}x_n} = \frac{qAD_p}{L_p} p_n \left(\exp\left(\frac{qV}{kT}\right) - 1 \right) \exp\left(-\frac{x_n}{L_p} \right) \quad (2.21\mathrm{d})$$

The electron and hole minority currents exponentially decay as x_p and x_n increase. This is because the number of minority carriers that have not recombined decreases as we move away from the junction. If we want to know the total injected current *before* any

recombination has taken place we must evaluate $I_n(x_p)$ at $x_p = 0$ and $I_p(x_n)$ at $x_n = 0$. This gives

$$I_n = -\frac{qAD_n}{L_n} n_p \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

$$I_p = \frac{qAD_p}{L_p} p_n \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

These results assume that carriers do not recombine while they cross over the depletion region. Carrier recombination or generation in the depletion region will be further discussed in Section 2.8.

Every time a minority carrier recombines, a majority carrier is required for the recombination event. As majority carriers are consumed by the minority carriers that are injected, additional majority carriers must be supplied by the external circuit. The flow of these majority carriers becomes the measured diode current. Since coordinates x_n and x_p point in opposite directions the total current I will be given by $I_p - I_n$ and we finally obtain the *diode equation*

$$I = qA \left(\frac{D_n}{L_n} n_p + \frac{D_p}{L_p} p_n \right) \left(\exp\left(\frac{qV}{kT}\right) - 1 \right) \tag{2.22a}$$

For large negative values of V, only the reverse saturation current flows and therefore

$$I_0 = qA \left(\frac{D_n}{L_n} n_p + \frac{D_p}{L_p} p_n \right) \tag{2.22b}$$

Figure 2.16 shows the excess carrier concentrations in the p-n junction for both forward and reverse bias. Note the exponential decay of excess carriers as distance from the depletion region boundaries increases.

The symmetrical minority hole and minority electron concentration profiles shown in Figure 2.16 imply equal n-type and p-type doping levels N_d and N_a as well as equal minority carrier lifetimes τ_n and τ_p. In practice, diodes often have $N_d \gg N_a$ or $N_d \ll N_a$, which means that very large ratios between I_n and I_p or very large ratios between hole and electron currents can occur.
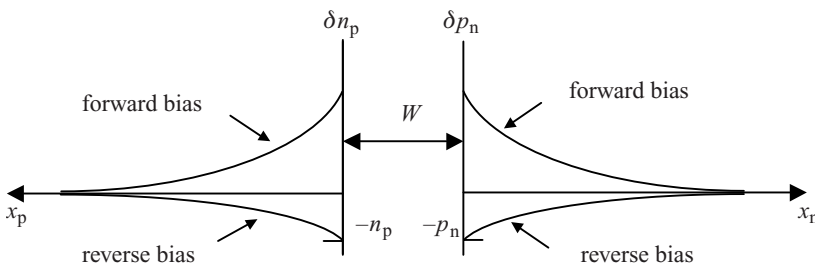
Figure 2.16 *Excess carrier concentration on either side of the p-n junction depletion region. For forward bias the excess concentration is positive and for reverse bias it is small and negative*

Another way to view the flow of current across a diode is to think in terms of minority carrier recombination. The area under the plot of δn_p versus x_p for the junction in Figure 2.16 is the total minority carrier charge Q_n injected into the p-side. From Equation 2.21,

$$Q_n = q A \int_0^\infty \delta n \left(x_p \right) dx_p = q A n_p \left(\exp \left(\frac{qV}{kT} \right) - 1 \right) \int_0^\infty \exp \left(-\frac{x_p}{L_n} \right) dx_p$$

$$= q A L_n n_p \left(\exp \left(\frac{qV}{kT} \right) - 1 \right)$$

and similarly,

$$Q_p = q A L_p p_n \left(\exp \left(\frac{qV}{kT} \right) - 1 \right)$$

Now, the total diode current may be determined. The diode current flow due to minority electrons will be $\frac{Q_n}{\tau_n}$ because the charge Q_n will recombine an average of once during recombination time τ_n and therefore charge Q_n must be resupplied by the external circuit. The corresponding current due to minority holes will be $\frac{Q_p}{\tau_p}$.

Total diode current is therefore

$$I = \frac{Q_n}{\tau_n} + \frac{Q_p}{\tau_p} = \left(\frac{q A L_n n_p}{\tau_n} + \frac{q A L_p p_n}{\tau_p} \right) \left(\exp \left(\frac{qV}{kT} \right) - 1 \right)$$

which is identical to Equation 2.22 since from Equation 1.46 $\frac{L_n}{\tau_n} = \frac{D_n}{L_a}$ and $\frac{L_p}{\tau_p} = \frac{D_p}{L_p}$.

The total diode current makes a transition from being composed of minority and majority carriers at the depletion region boundaries to being composed of majority carriers several diffusion lengths away from the depletion region on either side of the diode. From Equations 2.21 and 2.22 we have

$$I_n \left(x_n \right) = I - I_p \left(x_n \right) = q A \left[\frac{D_n}{L_n} n_p + \frac{D_p}{L_p} p_n \left(1 - \exp \left(-\frac{x_n}{L_p} \right) \right) \right] \left(\exp \left(\frac{qV}{kT} \right) - 1 \right)$$

At $x_n = 0$, $I_n \left(x_n \right)$ does not fall to zero. This is because an additional component of $I_n(x_n)$ must reach the depletion region to be injected to supply the p-side with its minority carriers. A similar expression results for $I_p \left(x_p \right)$. The resulting minority and majority currents are plotted as a function of x_p and x_n in Figure 2.17.
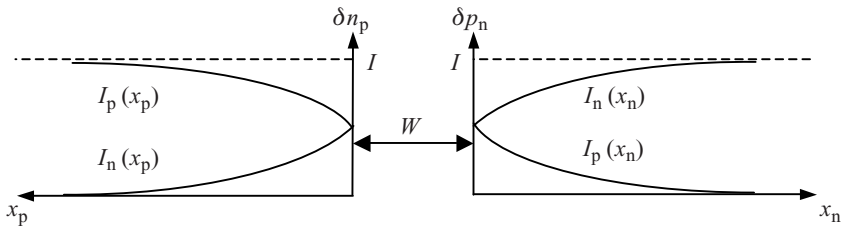
Figure 2.17   *Minority currents $I_n(x_p)$ and $I_p(x_n)$ as well as majority currents $I_p(x_p)$ and $I_n(x_n)$. The sum of the majority and minority currents is always the total current I. Each majority current is divided into two parts, one part supplying carriers to recombine with minority carriers and the other part being injected across the junction to supply the other side with minority carriers*

Example 2.4

(a) The silicon diode of Example 2.1 is forward biased and a current of 1 mA flows. Find the excess minority carrier concentration on either side of the junction at room temperature. Assume a carrier recombination time for electrons and holes in silicon of 2 μs and a junction area of 1 mm^2.

(b) Repeat for the p$^+$-n diode of Example 2.2b.

Solution

(a) From Appendix 2, for silicon $\mu_n = 1350\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$ and $\mu_p = 480\,\text{cm}^2\,\text{V}^{-1}$ s^{-1}. Therefore, using the Einstein relation,

$$D_n = \frac{kT}{q}\mu_n = 0.026\,\text{V} \times 1350\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1} = 3.51 \times 10^1\,\text{cm}^2\,\text{s}^{-1}$$

and

$$D_p = \frac{kT}{q}\mu_p = 0.026\,\text{V} \times 480\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1} = 1.25 \times 10^1\,\text{cm}^2\,\text{s}^{-1}$$

$$L_n = \sqrt{D_n\tau_n} = \sqrt{3.51 \times 10^1\,\text{cm}^2\,\text{s}^{-1} \times 2 \times 10^{-6}\,\text{s}} = 8.38 \times 10^{-3}\,\text{cm}$$

and

$$L_p = \sqrt{D_p\tau_p} = \sqrt{1.25 \times 10^1\,\text{cm}^2\,\text{s}^{-1} \times 2 \times 10^{-6}\,\text{s}} = 5.00 \times 10^{-3}\,\text{cm}$$

also,

$$n_p = \frac{n_i^2}{p_p} = \frac{\left(1.25 \times 10^{10}\,\text{cm}^{-3}\right)^2}{1 \times 10^{17}\,\text{cm}^{-3}} = 2.25 \times 10^3\,\text{cm}^{-3}$$

and

$$p_n = \frac{n_i^2}{n_n} = \frac{\left(1.5 \times 10^{10}\,\text{cm}^{-3}\right)^2}{1 \times 10^{17}\,\text{cm}^{-3}} = 2.25 \times 10^3\,\text{cm}^{-3}$$

From Equation 2.22 solving for V we obtain

$$V = \frac{kT}{q}\ln\left[\frac{I}{qA}\left(\frac{D_n}{L_n}n_p + \frac{D_p}{L_p}p_n\right)^{-1} + 1\right]$$

$$\frac{D_n}{L_n} = \frac{3.51 \times 10^1\,\text{cm}^2\,\text{s}^{-1}}{8.38 \times 10^{-3}\,\text{cm}} = 4.19 \times 10^3\,\text{cm}\,\text{s}^{-1}$$

$$\frac{D_p}{L_p} = \frac{1.25 \times 10^1\,\text{cm}^2\,\text{s}^{-1}}{5.00 \times 10^{-3}\,\text{cm}} = 2.50 \times 10^3\,\text{cm}\,\text{s}^{-1}$$

Now, $V = 0.026\,\text{V}\ln\left[\dfrac{10^{-3}\,\text{A}}{1.6 \times 10^{-19}\,\text{C} \times 10^{-2}\,\text{cm}^2}(4.19 \times 10^3\,\text{cm}\,\text{s}^{-1} \times 2.25\right.$

$$\left. \times 10^3\,\text{cm}^{-3} + 2.50 \times 10^3\,\text{cm}\,\text{s}^{-1} \times 2.25 \times 10^3\,\text{cm}^{-3})^{-1} + 1\right]$$

$$= 0.636\,\text{V}$$

Now on the n-side

$$Q_n = q\,AL_n n_p \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

$$= 1.6 \times 10^{-19}\,\text{C} \times 10^{-2}\,\text{cm}^2 \times 8.38 \times 10^{-3}\,\text{cm} \times 2.25$$

$$\times 10^3\,\text{cm}^{-3} \left(\exp\left(\frac{0.636\,\text{V}}{0.026\,\text{V}}\right) - 1 \right) = 1.27 \times 10^{-9}\,\text{C}$$

and on the p-side

$$Q_p = q\,AL_p p_n \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

$$= 1.6 \times 10^{-19}\,\text{C} \times 10^{-2}\,\text{cm}^2 \times 5.00 \times 10^{-3}\,\text{cm} \times 2.25$$

$$\times 10^3\,\text{cm}^{-3} \left(\exp\left(\frac{0.636\,\text{V}}{0.026\,\text{V}}\right) - 1 \right) = 7.56 \times 10^{-10}\,\text{C}$$

(b) For the p^+-n junction

$$n_p = \frac{n_i^2}{p_p} = \frac{(1.5 \times 10^{10}\,\text{cm}^{-3})^2}{1 \times 10^{18}\,\text{cm}^{-3}} = 2.25 \times 10^2\,\text{cm}^{-3}$$

and

$$p_n = \frac{n_i^2}{n_n} = \frac{(1.5 \times 10^{10}\,\text{cm}^{-3})^2}{1 \times 10^{16}\,\text{cm}^{-3}} = 2.25 \times 10^4\,\text{cm}^{-3}$$

Now on the n-side

$$Q_n = q\,AL_n n_p \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

$$= 1.6 \times 10^{-19}\,\text{C} \times 10^{-2}\,\text{cm}^2 \times 8.38 \times 10^{-3}\,\text{cm} \times 2.25$$

$$\times 10^2\,\text{cm}^{-3} \left(\exp\left(\frac{0.636\,\text{V}}{0.026\,\text{V}}\right) - 1 \right) = 1.27 \times 10^{-10}\,\text{C}$$

and on the p-side

$$Q_p = q\,AL_p p_n \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

$$= 1.6 \times 10^{-19}\,\text{C} \times 10^{-2}\,\text{cm}^2 \times 5.00 \times 10^{-3}\,\text{cm} \times 2.25$$

$$\times 10^4\,\text{cm}^{-3} \left(\exp\left(\frac{0.636\,\text{V}}{0.026\,\text{V}}\right) - 1 \right) = 7.56 \times 10^{-9}\,\text{C}$$

Note that holes injected from the heavily doped p-side into the lightly doped n-side dominate the current flow.

If a diode is very strongly forward biased the assumption of low-level injection will no longer be valid. This implies that majority carrier concentrations can no longer be taken as constant. Additional current flows now arise due to gradients in majority carrier
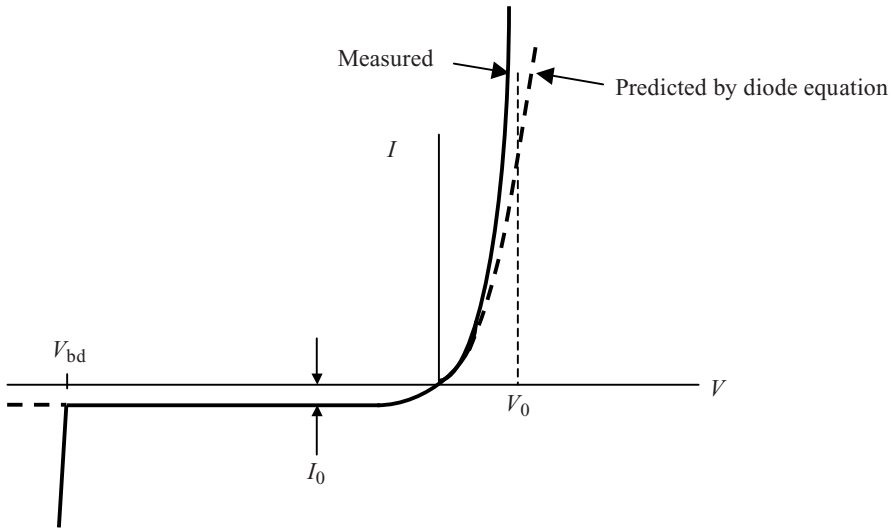
Figure 2.18   *Measured current–voltage characteristics of a diode as well as predicted characteristics based on the diode equation. A very steep increase in current as applied voltage approaches the built-in potential V_0 is observed in practice as well as an abrupt onset of reverse breakdown current at V_{bd}*

concentrations, and diode currents actually increase more rapidly with applied voltage predicted by the diode equation. This is illustrated in Figure 2.18 along with the predicted curve from the diode equation. Note that for forward bias the observed current–voltage characteristic rises very steeply as applied voltage V approaches V_0. This is consistent with the band model of Figure 2.5, which shows that the potential barrier height for majority carrier injection will approach zero as V approaches V_0. The reverse bias voltage V_{bd} will be discussed in Section 2.6.

A quantitative model that includes strong bias conditions can be derived but will not be undertaken in this book. Additional effects including bulk resistances of the neutral semiconductor regions and contact resistances must also be considered for an accurate model of measured diode characteristics.

2.6   Reverse Breakdown and the Zener Diode

When a large enough reverse bias is applied to a diode there will be an additional current flow, as shown for negative applied voltage in Figure 2.18. This current generally starts abruptly at a well-defined *reverse breakdown voltage V_{bd}* and must be carefully limited due to the large power $P = I V_{bd}$ that must be dissipated by the diode.

There are two possible mechanisms for this current flow. The first is *avalanche breakdown*. This may be understood from Figure 2.6, which shows the increased steepness of the band bending upon application of a reverse bias. This leads to a higher peak value of junction electric field ε_{bias} in the depletion region. If ε_{bias} exceeds the breakdown electric field
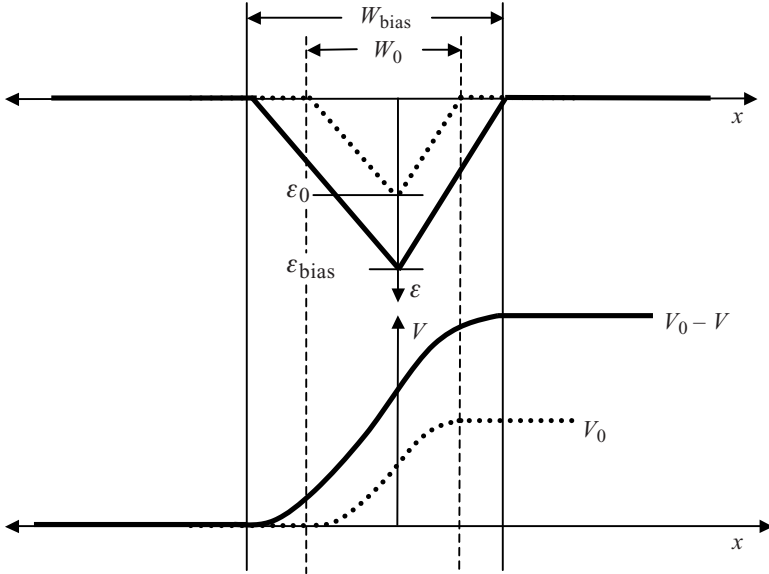
Figure 2.19 *Increase in depletion region width and increase in junction field with the application of a reverse bias for the p-n junction of Figure 2.13. The equilibrium conditions with depletion width W and peak electric field ε_0 are shown with dotted lines. With the application of reverse bias V (V negative) the depletion width increases to W_{bias} and the peak electric field increases to ε_{bias}*

strength for a given semiconductor then bound electrons that normally do not contribute to current flow may become available for conduction. This occurs through *field ionization* of semiconductor atoms. Once a small number of normally bound electrons is released these electrons can increase in number by *impact ionization* of other atoms. Since the ionization energy for atoms in a given semiconductor is very specific to that semiconductor material, this process occurs at a specific reverse voltage and then increases rapidly with a further increase in voltage.

If a negative bias voltage V is applied the potential barrier increases to $V_0 - V$. The depletion region width W_{bias} as well as ε_{bias} also increase, as shown in Figure 2.19. We can now modify the expressions we have for equilibrium conditions to calculate W_{bias} as well as ε_{bias}

The integral of electric field across the depletion region (Equation 2.13) becomes the area under the new ε versus x graph in Figure 2.19 and we obtain

$$V_0 - V = \frac{1}{2}bh = \frac{1}{2}W_{bias}\varepsilon_{bias}$$

From Equation 2.12 we can see that ε_{bias} increases linearly with an increase in depletion region width on either side of the junction, which means that both W_{bias} and ε_{peak} must be proportional to the square root of $(V_0 - V)$. Hence Equation 2.15a may be modified to

become

$$W_{bias} = \sqrt{\frac{2\epsilon_0\epsilon_r(V_0 - V)}{q}\left(\frac{1}{N_a} + \frac{1}{N_d}\right)} \qquad (2.23)$$

and from Equation 2.12 and Equation 2.15b or 2.15c we can write

$$\varepsilon_{bias} = -\frac{qW_{bias}}{\epsilon_0\epsilon_r}\frac{N_aN_d}{N_a + N_d} \qquad (2.24)$$

Thus in reverse bias, the depletion region width as well as peak electric field increase with increasingly negative bias voltage until ε_{bias} reaches the breakdown field of the semiconductor. At this point, no further increase in ε_{bias} can occur; however, electrons are generated at or near the junction, which are swept along by the electric field resulting in the reverse breakdown current.

From Equation 2.23 it is clear that the depletion width increases as the doping levels decrease. Since diodes often have much higher doping levels on one side than the other side we will consider a diode having $N_a \gg N_d$. From Equation 2.23 we can simplify the expression for depletion region width to obtain

$$W_{bias} = \sqrt{\frac{2\epsilon_0\epsilon_r(V_0 - V)}{qN_d}} \qquad (2.25)$$

We can now use Equation 2.24 to obtain the peak electric field as

$$\varepsilon_{bias} = -\sqrt{\frac{2qN_d(V_0 - V)}{\epsilon_0\epsilon_r}} \qquad (2.26)$$

Hence the magnitude of the breakdown voltage may be increased by *decreasing* the doping level on at least one side of a diode. This is easy to visualize, since a decreased doping level leads to a wider depletion region with lower electric fields present for a given bias voltage V.

Another mechanism may also cause reverse breakdown current. This occurs in p-n junctions that have simultaneously high values of both n-type and p-type doping. From Equation 2.15a, if both N_a and N_d are large then depletion width W will be small. This is illustrated in Figure 2.20. W may be small enough to allow tunnelling of electrons directly from the valence band on the p-side into the conduction band on the n-side. This differs from avalanche breakdown since neither field-ionization nor impact-ionization occur.

This tunnelling type of breakdown is properly referred to as *Zener breakdown* although the *Zener diode* has come to mean a diode used specifically for its reverse breakdown characteristics whether caused by avalanching or tunnelling processes. In practice, diodes with small breakdown voltages of a few volts involve tunnelling and diodes with higher breakdown voltages of 5 volts or more involve avalanching. There are often combinations of these two mechanisms occurring simultaneously in diodes with intermediate breakdown voltages.
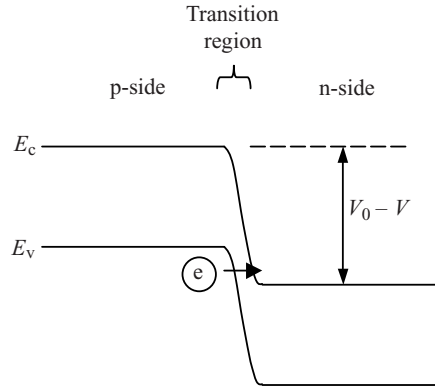
Figure 2.20   *Tunnelling of valence-band electron from valence band on p-side to conduction band on n-side upon application of a small reverse bias voltage. Note that there is a large supply of valence band electrons on the p-side. In comparison there is only a small supply of thermally generated minority carrier electrons that result in current I_0. This explains how the reverse current can be much larger than I_0 as shown in Figure 2.18, when V exceeds the breakdown voltage V_{bd}*

2.7   Tunnel Diodes

An important extension of the tunnelling mechanism that operates in Zener breakdown occurs in the *tunnel diode*. If the doping levels are further increased to become very high on both the n-side and the p-side of a p-n junction it becomes possible to have tunnelling currents flowing in both directions rather than in only one direction. This is illustrated in Figure 2.21. The doping, known as *degenerate* doping, is now high enough to push the Fermi energy into the conduction band on the n-side and into the valence band on the p-side yielding high electron and hole concentrations as well as the alignment of electron energy states. In practice this occurs in silicon at doping concentrations of over approximately $1 \times 10^{19} \, \mathrm{cm}^{-3}$.
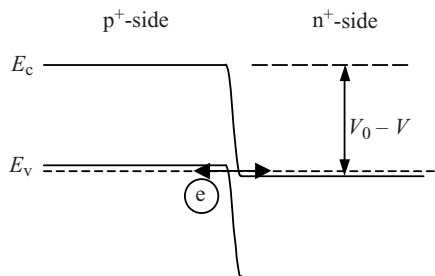
Figure 2.21   *In a tunnel diode the depletion width is very narrow due to the use of degenerate p^+ and n^+ doping. In addition to the narrow depletion region, the Fermi level enters the bands on either side of the diode resulting in the alignment of electron energy states in the conduction band on the n-side with valence electron states on the p-side. Electron tunnelling occurs in either direction as shown*
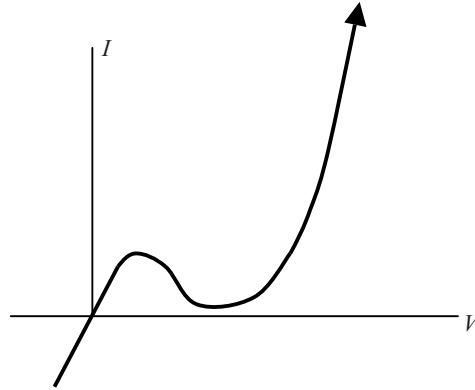
Figure 2.22 *Current–voltage (I–V) characteristic of a tunnel diode. At low voltages, tunnelling currents result in significant current flow in both directions. At higher positive bias voltages, electrons in the conduction band on the n-side will no longer be aligned with the valence band on the p-side. This will prevent tunnelling and current flow will therefore decrease. Current flow will eventually rise upon further increase of forward bias since the potential barrier decreases as in a normal p-n junction*

As in the Zener diode, electrons in the valence band of the p-type material can tunnel directly into the conduction band of the n-type material. In addition, however, electrons in the conduction band of the n-type material can now tunnel directly into unfilled energy states on the p-side. At the Fermi energy, half the electron states are vacant in the valence band on the p-side. Since current now flows in either direction, the p-n junction no longer behaves like a diode. Even at the lowest applied voltages of either polarity, tunnelling current flows. At higher voltages, however, normal diode behaviour is regained because the alignment between electron states on either side of the junction is lost. The resulting *I-V* characteristic is shown in Figure 2.22.

The tunnel diode is of particular relevance to high performance solar cell devices that require a series connection between two or more p-n junctions. If a pair of p-n junctions, each consisting of an n-type layer and a p-type layer, are to be stacked on top of each other in a series-connected arrangement an unwanted n-p junction will exist between the two p-n junctions. If the series-connected pair of junctions is connected to a voltage source arranged to apply a forward bias to the junctions, the unwanted n-p junction will be reverse biased, effectively limiting the forward current that can flow. Provided, however, that the unwanted n-p junction is designed to be a tunnel diode by appropriately doping its n- and p-layers heavily, its influence on the overall forward current can be minimized. This will be further discussed in Chapter 4, and can be extended to three or more series-connected p-n junctions.

2.8 Generation/Recombination Currents

We now need to develop a more realistic picture of carrier behaviour within the depletion region. Until now, we have specifically neglected carrier recombination or generation in this
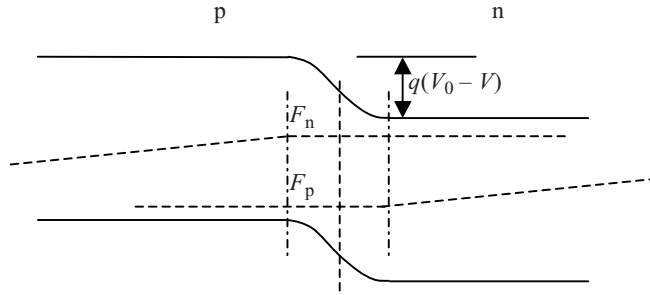
Figure 2.23 *The quasi-Fermi levels within the depletion region are shown. Although the depletion region is created by the recombination of charges in equilibrium, once injection takes place in forward bias, excess carriers must flow through this region*

region; however, for both LEDs and solar cells these processes are important, particularly carrier recombination under forward bias conditions.

In forward bias, excess carriers *must* actually be present in the depletion region as they are injected across it. In Figure 2.15a we showed the quasi-Fermi levels for a forward-biased p-n junction. The depletion region is shown again in Figure 2.23. Note that the hole and electron concentrations due to excess carriers are not zero in the depletion region when forward bias is applied.

In the depletion region, there is an increased likelihood that these excess carriers will recombine through traps. This is because only excess carriers exist here and there are no normally occurring majority carriers available to participate in the recombination process. Direct electron-hole pair recombination is proportional to the product np as discussed in Section 1.15 without considering traps.

In Section 1.19 we argued that assuming the existence of traps at approximately mid-gap is justifiable due to defects in the semiconductor crystal. We will therefore assume a trap energy E_t at mid-gap that dominates the recombination process.

Neglecting direct electron-hole pair recombination in the depletion region and considering only trapping of carriers we can use Equation 1.49 for the hole recombination rate with a fixed trap concentration and with $G_{op} = 0$. We will use this expression but simply assume a fixed trap concentration in place of a fixed electron concentration and use carrier lifetimes τ_n and τ_p that result from trapping. Hence, for holes, in the absence of carrier generation,

$$\frac{d\delta p(t)}{dt} = -\frac{\delta p}{\tau_p}. \tag{2.27a}$$

and if the minority carriers were electrons,

$$\frac{d\delta n(t)}{dt} = -\frac{\delta n}{\tau_n}. \tag{2.27b}$$

From Equation 1.60a and Figure 2.15a,

$$\delta n \delta p = n_0 p_0 \exp\left(\frac{F_n - F_p}{kT}\right) = n_i^2 \exp\left(\frac{qV}{kT}\right) \tag{2.28}$$

If the quasi-Fermi levels are approximately symmetric about mid-gap then $\delta n = \delta p$. Since trap energy $E_t \cong E_f$, we should consider that half the traps are filled and half are empty. This reduces the recombination rate by a factor of two. From Equations 2.28 and 2.27 we obtain for holes

$$\frac{d\delta p(t)}{dt} = -\frac{n_i}{2\tau_p} \exp\left(\frac{qV}{2kT}\right)$$

and for electrons,

$$\frac{d\delta n(t)}{dt} = -\frac{n_i}{2\tau_n} \exp\left(\frac{qV}{2kT}\right)$$

A recombination rate R for holes recombining through traps may now be defined as

$$R = -\frac{d\delta p(t)}{dt} = \frac{n_i}{2\tau_p} \exp\left(\frac{qV}{2kT}\right)$$

and for electrons,

$$R = -\frac{d\delta n(t)}{dt} = \frac{n_i}{2\tau_n} \exp\left(\frac{qV}{2kT}\right)$$

Since carriers that recombine must be replaced by current flowing through the external circuit, we can regard a part of the total diode current as being proportional to R. As we move away from the junction and reach the edges of the depletion region the recombination rates will return to the rates we calculated to obtain Equation 2.22, yielding diode current proportional to $\exp(\frac{qV}{kT})$. The overall diode current including trapping therefore varies with V as

$$I \propto \exp\left(\frac{qV}{nkT}\right)$$

where n is the *ideality factor*, which varies between 1 and 2. The precise value of n depends on the ratio of the carrier lifetime in the absence of traps to the corresponding lifetime with the traps present. In addition, n depends on the effective depletion region width and it also depends on the magnitude of the current. At high diode currents, n approaches 1 since trapping will become less significant than at low diode currents.

Carrier generation within the depletion region can also be important in diodes that are reverse biased or weakly forward biased. If trapping sites capture electrons and holes then these sites can also generate these carriers by thermal activation. Normally trap generation is in thermal equilibrium with trap recombination. Under reverse bias or weak forward-bias conditions, however, carriers generated by traps may be accelerated away from the depletion region before they can recombine, which will increase the reverse saturation current. This phenomenon can be observed when reverse saturation current I_0 increases with reverse bias rather than remaining fixed. If carriers released from traps in the depletion region dominate I_0 and these traps are uniformly distributed over the depletion region, then I_0 would increase with approximately the square root of the magnitude of reverse-bias voltage. See Equation 2.25, which shows that depletion width increases as the square root of voltage for $V \gg V_0$. In practical silicon diodes at room temperature, thermal generation current within the depletion region often dominates the reverse saturation current.

If light is incident on the depletion region carrier generation can occur. This is relevant to solar cells and photodiodes and will be discussed in detail in Chapter 4.

2.9   Ohmic Contacts, Schottky Barriers and Schottky Diodes

Without electrical contacts, semiconductor devices are usually not useful. The flow of electrons to and from metals that connect semiconductor devices to external circuitry is dependent on a variety of influences. We can model the most important mechanisms at play and we can use the band model to explain the current flow. In addition we need to consider the factors that influence current flow.

A metal has a partly filled conduction band, which means that the Fermi energy in the metal lies within this band. If no charges are exchanged between the metal and the semiconductor, we can draw an idealized metal-semiconductor contact in equilibrium by maintaining a constant Fermi level for the metal-semiconductor junction as in Figure 2.24. A small barrier E_b for electrons appears at the junction where $E_b = E_c - E_f$. In this diagram electrons can easily flow from the semiconductor into the metal; however, electrons from the metal need to overcome energy barrier E_b to flow into the semiconductor. Mobile charge can flow across the metal-semiconductor contact, which may either deplete or accumulate near the semiconductor surface and therefore the conditions of Figure 2.24 do not correctly represent equilibrium conditions.

Provided the doping is high enough, $E_c - E_f$ can be made small enough to permit efficient electron flow in both directions because E_f moves closer to the conduction band edge as doping increases, and an *ohmic contact* is formed. In some cases it is of interest to have diode behaviour, but in many cases the desire is for ohmic behaviour in which current flow is enabled and ohmic contacts are achieved, which allow for current flow in either direction with minimal voltage drop across the junction.
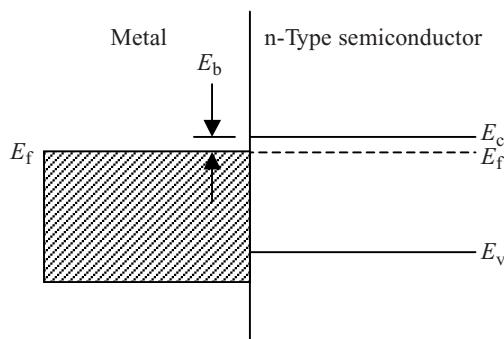
Figure 2.24   *Metal-semiconductor contact for an n-type semiconductor without any flow of charge between the two sides. The predicted barrier height $E_b = E_c - E_f$ and the flat bands shown are not achieved in real devices due to charge flow and charges at the metal-semiconductor interface that cause band bending and associated electric fields at the interface and in the semiconductor*

Real metal-semiconductor contacts generally result in band-bending due to charges that flow across the metal-semiconductor interface. In addition charges from various sources may accumulate at or near the metal-semiconductor interface. A number of sources of charge at or near the junction exist:

1. Electron flow between the metal and semiconductor. Electron affinities of the semiconductor atoms and the metal atoms do not generally match. This is quantified at a free surface by a workfunction difference between the two materials. Since we are interested in an interface and workfunctions are properly measured at a surface in a vacuum, the contribution of this workfunction difference to the metal-semiconductor junction is understood by using *modified workfunctions* that take the actual interface bonding conditions into account. As a result, electrons will flow between the metal and semiconductor to minimize the potential energy of the system. The result will be a net charge in the semiconductor and a net charge in the metal. This charge will be significant within one diffusion length from the junction in the semiconductor.
2. Dangling bonds at the semiconductor surface. These form since the lattices of the metal and semiconductor do not match.
3. Dangling metal bonds at the metal surface. These form since the lattices of the metal and semiconductor do not match. In addition, metals are generally polycrystalline and there will be many crystallographic orientations to consider in one device.
4. Doping effects from metal atoms that diffuse into the semiconductor. If the metal atoms have a different valence number from the semiconductor atom then a charge will be associated with each metal ion in the semiconductor.
5. The *Schottky effect* lowers the effective barrier height due to a coulomb attraction between charges in the semiconductor and the metal, which is not included in Figure 2.24. This will not be discussed in detail in this book.
6. Oxygen or other impurity incorporation at the metal/semiconductor surfaces near the interface. For example, it is very difficult to completely eliminate oxygen at a metal-semiconductor interface since most metals and common semiconductors such as silicon react strongly with oxygen. Oxide layers tend to trap positive charges due to electron loss from metal/semiconductor dangling bonds within the oxide layer.

The result is that it is virtually impossible to predict the precise amount and polarity of charge present near metal-semiconductor interfaces. This charge will result in band bending, which has a strong effect on the electrical properties of the metal-semiconductor junction. In practice, to achieve specific electrical properties, there are known recipes for metal-semiconductor junctions that yield good results for a variety of applications.

We will consider two cases and then see how these cases affect the behaviour of the metal-semiconductor contact using an n-type semiconductor. In Figure 2.25a positive charge is trapped at the semiconductor surface. In principle this positive charge could be a result of several of the sources mentioned above. This causes the Fermi energy to move closer to the conduction band in the semiconductor near the interface because the trapped positive charge attracts electrons in the conduction band towards the interface. The Fermi energy may actually enter the conduction band, which means that there is enough positive charge present at the semiconductor surface to attract conduction band electrons to fill the low-lying conduction band states with more than a 50% probability. This results in an ohmic contact.

Metal   n-type semiconductor

(a)

(b)
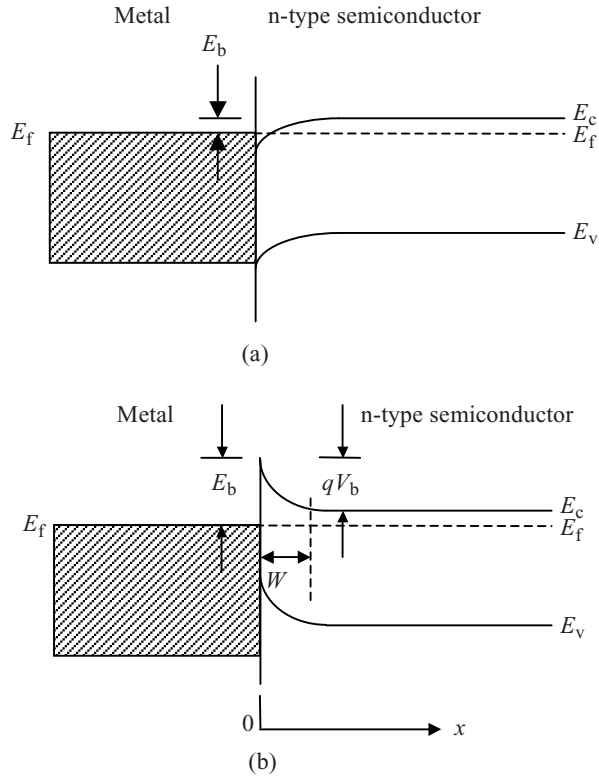
Figure 2.25 *Metal-semiconductor contact energy band diagrams under various conditions. The subscripts 'b' refer to the fact that these are 'built-in' and are present without the application of an external voltage. (a) If the interface is positively charged then the band-bending will be as shown. This forms an ohmic contact provided that the semiconductor doping level is high enough to make the energy barrier E_b small. (b) If the interface is negatively charged then band bending will result in a large energy barrier E_b, which blocks electron flow from the metal to the semiconductor, as well as a depletion region in the semiconductor. A Schottky diode is formed*

The formation of a *Schottky diode is* illustrated in Figure 2.25b, in which negative charge is trapped at the junction. The Schottky diode is also referred to as an *MS diode* or *Metal-Semiconductor diode*. This diode will now be discussed and modelled. It has the advantage of being a diode that can be fabricated at low cost since a metal thin film may be deposited on a semiconductor surface by a variety of well-developed vacuum deposition techniques such as evaporation or sputtering. The evaporation of an aluminium thin film on the surface of an n-type silicon semiconductor moderately doped n-type to $n \cong 10^{16}$ cm^{-3} is an example of a Schottky diode structure. The diode characteristics will depend on various details of the processing conditions including the crystallographic orientation of the silicon wafer, the degree of trapped oxygen at the interface between the aluminium and the silicon, and the cleaning procedure used to prepare the silicon surface prior to deposition; however,

a resulting band diagram similar to that shown in Figure 2.25b is routinely attainable in production.

In Figure 2.25b, the negative charge trapped at the semiconductor surface is a common situation due to dangling bonds that form at a semiconductor surface. This was discussed in Section 1.19. Conduction band electrons are repelled away from the negative trapped charge leaving a depletion region in the semiconductor near the interface. The Fermi level gets pinned at approximately mid-gap at the semiconductor surface. A large barrier for electron flow from the metal to the semiconductor forms, which blocks current flow even when an external voltage is applied. The barrier for electrons leaving the semiconductor is labelled qV_b and is formed due to the band bending. If a forward bias voltage V is applied across the metal-semiconductor junction such that the metal is positive then this barrier will be reduced to $q(V_b - V)$, allowing current flow. If a reverse bias voltage of magnitude V is applied to the metal relative to the semiconductor the height of the barrier qV_b will be increased to $q(V_b + V)$ and current flow will be blocked. Diode behaviour will therefore be obtained.

We can apply the depletion approximation to the Schottky diode to simplify the treatment of the depletion region. If we assume that the depletion region is fully depleted over a width W then the depletion region contains only static charges due to ionized donors. If the semiconductor is n-type with a uniform doping concentration N_d then the total charge in the depletion region in units of coulombs per unit junction area is qN_dW. This charge determines the electric field at the semiconductor surface since we can apply a Gaussian surface to the diode that encloses the entire depletion region. Now, using Gauss's law from Equation 2.11 and assuming a junction area A we have

$$\oint_S \vec{\varepsilon} \cdot \overrightarrow{ds} = \frac{Q}{\epsilon} = \frac{qN_dWA}{\epsilon} \tag{2.29}$$

The electric field in the neutral region of the semiconductor is zero. Provided the junction cross-section dimensions are much larger that W we can conclude from Equation 2.29 that

$$\varepsilon_s = \frac{qN_dW}{\epsilon} \tag{2.30}$$

where ε_s is the magnitude of the electric field at the semiconductor surface where $x = 0$. Using Gaussian surfaces that enclose smaller portions of the depletion region the electric field as a function of depth x may be found (see Problem 2.5) and as a result the magnitude of the electric field as a function of depth $\varepsilon(x)$ over the range $0 \leq x \leq W$ is found to be

$$\varepsilon(x) = \frac{qN_dW}{\epsilon} \frac{(W-x)}{W} \tag{2.31}$$

Note that if $x = 0$ then from Equations 2.30 and 2.31 $\varepsilon(x) = \varepsilon_s$. $\varepsilon(x)$ falls linearly as a function of x until it reaches zero at $x = W$. This is similar to a p$^+$-n junction. See Example 2.2.

The potential difference arising from this electric field over the depletion region may be obtained by integrating Equation 2.31, and we therefore have

$$V_b = -\int_0^W \varepsilon(x)dx = \frac{qN_dA}{\epsilon} \int_0^W (W-x)dx = \frac{qN_d}{\epsilon} \frac{W^2}{2}$$

Solving for W we obtain

$$W = \sqrt{\frac{2 \epsilon V_b}{q N_d}}$$

If a reverse bias voltage V is applied, this voltage will add to the built-in voltage and the depletion width will increase to become

$$W = \sqrt{\frac{2 \epsilon (V_b + V)}{q N_d}}$$

The current flow through a Schottky diode differs from a normal semiconductor p-n junction because majority carriers dominate the process. In the Schottky diode of Figure 2.25b, n-type electron current flow occurs in both the metal and the semiconductor, and we need to consider the flow of electrons across a potential barrier, either from the metal to the semiconductor, or from the semiconductor to the metal. The net current flow will now be the difference between the electron flow in these two directions. We will neglect hole current in the semiconductor, which is small.

Electrons in the metal can overcome a potential barrier by a process called *thermionic emission*. Electrons occupy a range of energy states in the metal and only a fraction of these electrons have enough kinetic energy to overcome the relevant energy barrier at the junction. The fraction of electrons able to overcome the barrier is determined by temperature T as well as the height of the energy barrier E_b.

Consider the Schottky diode in equilibrium in Figure 2.25b. If electron current I_e flows by thermionic emission from the metal into the semiconductor then in order to satisfy equilibrium conditions the *net* current flow must be zero. We can therefore conclude that an equal and opposite electron current flows from the semiconductor into the metal. This latter current flows in the same way as current flows in a p-n junction. The electrons in a Schottky diode flowing from the n-type semiconductor towards the metal experience an energy band shape within the n-type semiconductor that is the same as in the n-type semiconductor of a p^+-n junction and will therefore follow a current–voltage relationship identical to that obtained in a p^+-n junction, which we described in Example 2.2. Hence

$$I = I_e \left(\exp\left(\frac{qV}{kT}\right) - 1 \right) \qquad (2.32)$$

Here we interpret I_e to be the current flowing by thermionic emission from the metal to the semiconductor over energy barrier E_b at temperature T. The form of Equation 2.32 satisfies the condition that for large negative values of V only the thermionic emission current I_e flows and if $V = 0$ then $I = 0$. Note that the mechanism for the flow of I_e is very different from the origin of saturation current I_0 in a normal p-n junction, which is due to minority carrier drift.

The determination of I_e is based on the fundamental behaviour of a metal-vacuum interface. Consider a metal-vacuum interface of area A at temperature T where the metal workfunction is Φ_m. Electrons with enough kinetic energy may overcome the workfunction, leaving the metal surface and entering the vacuum. We assume that the electrons in the vacuum cannot flow away unless they return to the metal by again entering the metal surface.
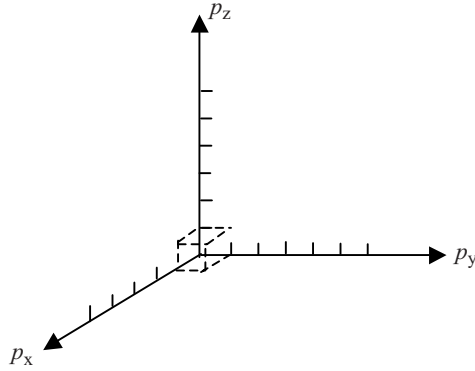
Figure 2.26 *Momentum space is equivalent to reciprocal space of Figure 1.11, but each axis of Figure 1.11 is multiplied by* h *to convert the space to momentum space in which the x-axis is marked in momentum units of $\frac{h}{a}$, the y-axis in momentum units of $\frac{h}{b}$, and the z-axis in momentum units of $\frac{h}{c}$. This is the model we use for the electrons in the vacuum adjacent to the metal surface*

This means that an equilibrium balance between electrons in the metal and electrons in the vacuum will be reached.

The electrons in the metal occupy energy levels that depend upon the band structure in the metal; however, the electrons in the vacuum can be treated more readily since they are free electrons inside a volume that encloses the vacuum. If we define the Fermi energy in the metal to be zero on our relative energy scale then the electrons in the vacuum have a potential energy equal to Φ_m in addition to their kinetic energy. These electrons can be treated as in Section 1.9, where we derived the density of states function for electrons in an infinite-walled potential box. In Figure 1.11 a reciprocal space is shown in which reciprocal lattice points represent allowed solutions to Schrödinger's equation for electrons in an infinite-walled potential box. This reciprocal space may be redrawn as a momentum space, as shown in Figure 2.26, in which each axis from Figure 1.11 is multiplied by a factor of h. Note that $p = hk = \frac{h}{\lambda}$.

In momentum space, the points in reciprocal space in Figure 1.11

$$\left(\frac{n_x}{a}, \frac{n_y}{b}, \frac{n_z}{c} \right)$$

correspond in momentum space to the points*

$$\left(\frac{n_x h}{a}, \frac{n_y h}{b}, \frac{n_z h}{c} \right)$$

---

*The attentive reader will notice that based on Figure 1.11 $p = \frac{hn_x}{2a}$. The factor of 2 is omitted in the denominator because the boundary condition applied to Equation 1.27 is altered for *travelling* electron waves rather than standing waves. The relevant boundary condition applied to Equation 1.27 becomes $k_x = \frac{2n_x\pi}{a}$ for travelling electron waves and hence $p = \frac{h}{2\pi}k = \frac{h}{2\pi} \times \frac{2n_x\pi}{a} = \frac{hn_x}{a}$. See Problem 2.12.

Each point represents two allowed electron states including spin, as explained in Section 1.9. It follows therefore that the number of electrons per unit crystal volume occupying the states in the range of momentum values dp_x, dp_y, dp_z is

$$dn = \frac{2}{h^3} dp_x dp_y dp_z F(E)$$

where the Fermi–Dirac function $F(E)$ is the probability of occupancy by an electron of any given state in the vacuum. We need to be careful, however, to include the potential energy of the vacuum. Since any electron in the vacuum has a potential energy term Φ_m as well as kinetic energy term $\frac{p^2}{2m}$, the form of the Fermi–Dirac function of Equation 1.23 should be

$$F(E) = \frac{1}{1 + \exp\left(\dfrac{E + \Phi_m}{kT}\right)}$$

Now if the metal-vacuum interface lies in a plane normal to the x-axis then the vacuum electrons incident on the interface will have a momentum in the x-direction and the resulting current flow across the interface due to electrons dn will be

$$dJ_x = qv_x dn = \frac{2qp_x}{mh^3} f(E) dp_x dp_y dp_z$$

since $p_x = mv_x$.

The total current density J_x that can flow across the metal-vacuum interface is now obtained by integrating over all ranges of electron energy. It is possible to simplify the integral since the energy ranges in the Fermi–Dirac function are much larger than kT and therefore the denominator of $F(E)$ is large. We are now justified in replacing the Fermi–Dirac distribution with the Boltzmann function and we have

$$F(E) \cong \frac{1}{\exp\left(\dfrac{E + \Phi_m}{kT}\right)} = \exp\left(-\frac{E + \Phi_m}{kT}\right)$$

where

$$E = \frac{p_x^2 + p_y^2 + p_z^2}{2m}$$

and therefore

$$F\left(p_x, p_y, p_z\right) \cong \exp\left(-\frac{p_x^2 + p_y^2 + p_z^2}{2mkT} - \frac{\Phi_m}{kT}\right)$$

Now,

$$J_x = \frac{2q}{mh^3} \exp\left(-\frac{\Phi_m}{kT}\right) \iiint p_x \exp\left(-\frac{p_x^2 + p_y^2 + p_z^2}{2mkT}\right) dp_x dp_y dp_z$$

$$= \frac{2q}{mh^3} \exp\left(-\frac{\Phi_m}{kT}\right) \int_0^\infty p_x \exp\left(-\frac{p_x^2}{2mkT}\right) dp_x \int_{-\infty}^\infty \exp\left(-\frac{p_y^2}{2mkT}\right) dp_y$$

$$\times \int_{-\infty}^\infty \exp\left(-\frac{p_z^2}{2mkT}\right) dp_z$$

The limits on the x-axis integral are between 0 and ∞ since only electrons in vacuum travelling in one direction along the x-axis will encounter the metal surface. The limits on the y-axis and z-axis integrals are between $-\infty$ and ∞ since we want to include all momentum components in the y and z directions.

Using the following standard integrals

$$\int_{-\infty}^{\infty} \exp\left(-ax^2\right) dx = \sqrt{\frac{\pi}{a}}$$

and

$$\int_{0}^{\infty} x \exp\left(-ax^2\right) dx = \frac{1}{2a}$$

we obtain

$$J_x = \frac{2q}{mh^3}\pi(2mkT)\exp\left(-\frac{\Phi_m}{kT}\right)\int_{0}^{\infty} p_x \exp\left(-\frac{p_x^2}{2mkT}\right) dp_x$$

$$= \frac{qm}{2\pi^2\hbar^3}(kT)^2 \exp\left(-\frac{\Phi_m}{kT}\right)$$

which in thermal equilibrium gives us a current density from the metal to the vacuum of

$$J_x = -\frac{qm}{2\pi^2\hbar^3}(kT)^2 \exp\left(-\frac{\Phi_m}{kT}\right)$$

and the magnitude of the current I_e leaving the metal surface of area A is

$$I_e = -\frac{qmA}{2\pi^2\hbar^3}(kT)^2 \exp\left(-\frac{\Phi_m}{kT}\right) \tag{2.33}$$

Equation 2.33 is known as the *Richardson–Dushman equation for thermionic emission*. It is important to remember that this current is an equilibrium current at a temperature T. This means that if we apply an electric field to draw off the electrons we are not truly measuring the equilibrium thermionic emission current; however, the measurement will approach the equilibrium current if the applied electric field is small.

If the vacuum is now replaced by silicon then the work function Φ_m is replaced by E_b and the electron mass m is replaced by the effective electron mass in the semiconductor m_e^*. Now Equation 2.33 becomes

$$I_e = \frac{(kT)^2 m_e^* q A}{2\pi^2\hbar^3} \exp\left(-\frac{E_b}{kT}\right) \tag{2.34}$$

Finally the Schottky diode current–voltage relationship is obtained from Equations 2.32 and 2.34 as

$$I = \frac{(kT)^2 m_e^* q A}{2\pi^2\hbar^3} \exp\left(-\frac{E_b}{kT}\right)\left(\exp\left(\frac{qV}{kT}\right) - 1\right)$$

The Schottky diode current increases exponentially as the barrier height E_b decreases. The current increases as a function of increasing temperature. Current dependence on applied voltage approaches an exponential function as forward bias increases.
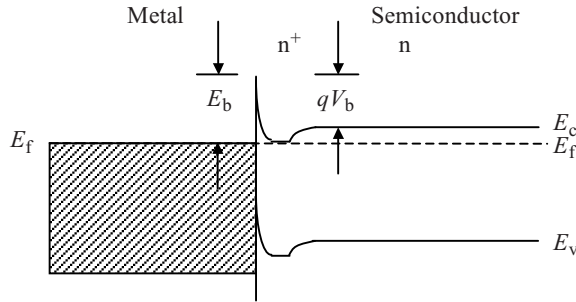
Figure 2.27 *If the interface is negatively charged and the semiconductor is strongly doped to form an n^+ region near the interface then band bending can result in a very narrow energy barrier of height E_b, which permits electron flow by tunnelling through the barrier and an ohmic contact is formed, as well as an additional built-in barrier qV_b formed due to band bending in the semiconductor*

The Schottky diode is not normally used directly as a solar cell or an LED; however, the concept of thermionic emission is important and is applicable to abrupt interfaces that form abrupt changes in potential energy in devices (see Section 2.10). In addition, organic devices discussed in Chapter 6 often require the application of thermionic emission (see Section 6.3).

We will now return to the formation of ohmic contacts on semiconductors, in which case the current–voltage characteristic of a Schottky diode must be avoided. There is a way to form an ohmic contact even if a negative charge forms at the interface of the semiconductor and metal, which would normally form a Schottky diode. Provided the semiconductor is very highly doped near the semiconductor surface, the width of the resulting depletion region can be made small enough to permit electron tunnelling directly from the metal to the semiconductor conduction band. This is illustrated in Figure 2.27. The elimination of an energy barrier is therefore not a prerequisite for an effective ohmic contact, and the width of the barrier determines its behaviour.

An analogous set of junctions can form in the case of a metal-semiconductor junction with a p-type semiconductor. We show one example of a p-type semiconductor-metal junction in Figure 2.28, which is equivalent to Figure 2.25a for the n-type case. Here the energy barrier is small and ohmic behaviour results; however, since current in the p-type semiconductor is predominantly hole current, electrons flowing from the metal into the semiconductor recombine with holes and allow more holes to flow towards the interface. Although the directions of hole and electron flows are opposite to each other, they constitute the same current flow. This is an effective ohmic contact and a p+ region is formed near the interface.

In practice an effective ohmic contact on n-type silicon may be formed by diffusing an extra n-type dopant such as phosphorus or antimony into the silicon surface forming an n+ region before a metal such as aluminium is applied resulting in a tunnelling ohmic contact similar to Figure 2.27. For p-type silicon aluminium may also be used. After depositing aluminum a brief heat treatment can be used to diffuse the aluminium into the silicon forming both a p+ region and the metal contact. In this way aluminium can serve as metallization for both n-type and p-type ohmic contacts on silicon.
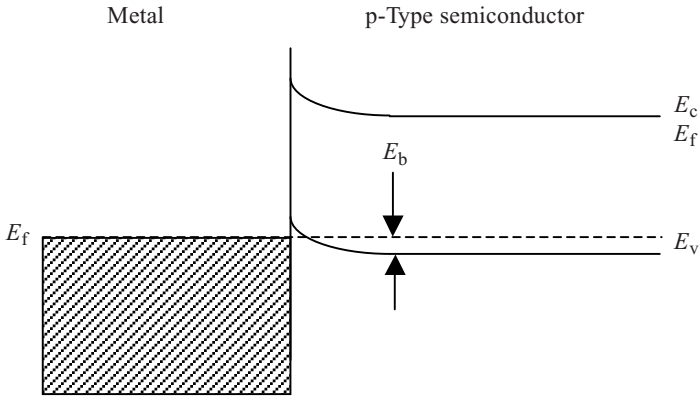
Figure 2.28   *Example of an ohmic contact between p-type silicon and a metal. Electrons from the metal recombine with a high concentration of holes that accumulate near the surface of the p-type semiconductor. Each hole that recombines allows another hole to take its place, resulting in continuous current flow*

In practice the formation of effective ohmic contacts to any given semiconductor is experimentally determined and optimized and cannot be fully modelled theoretically. A wide range of metals, dopants and heat treatments are known that form effective ohmic contacts on a wide range of semiconductors.

2.10   Heterojunctions

All the p-n junctions we have described until now make use of a single semiconductor material that is doped to form the n and p regions. The dopant concentration is typically in the parts per million range, which means that the semiconductor is really based on a single element in the case of elemental semiconductors such as Si, or a single compound in the case of compound semiconductors such as GaN.

If a p-n junction is formed from two different semiconductors, one on either side of the junction, then a *heterojunction* is formed. An example of this might be p-type GaAs on one side of the junction and n-type $Ga_{0.8}Al_{0.2}As$ on the other side. Since there is only a very small change in lattice constant for this system, as shown in Figure 1.30, a high-quality single-crystal p-n junction may be achieved. A band diagram for this type of junction is shown in Figure 2.29. Note that there is band bending near the junction due to the formation of a depletion region. In addition conduction and valence band offsets ΔE_c and ΔE_v arise. These band offsets imply highly localized electric fields and corresponding charge densities in a very thin sheet at the junction. As in the case of metal-semiconductor junctions, interface charge can arise from several sources. For a single-crystal semiconductor the dominant sources are differing electron affinities between the two semiconductors and anomalies in the bonding at the interface that give rise to localized polarization charges. In addition, since lattice constants are usually not perfectly matched, the inevitable stress and strain induced will further contribute to this local charge. Should dislocations develop, additional charges could exist.
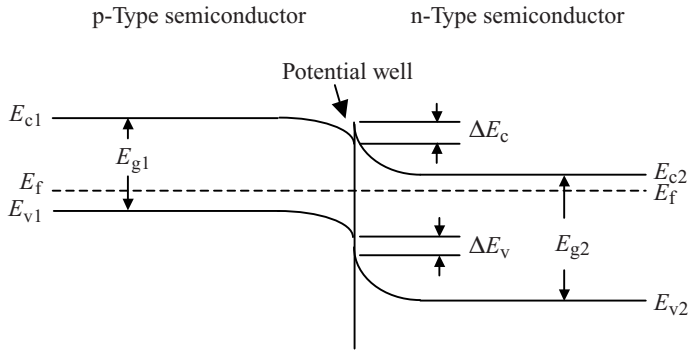
Figure 2.29   *Example of heterojunction formed between p-type GaAs and n-type Ga$_{1-x}$Al$_x$As*

There is no reason for ΔE_c and ΔE_v to be equal. Their relative values are determined by the interface charge; however, in all cases the relationship $\Delta E_c + \Delta E_v = \Delta E_g$ must hold. One obvious consequence of the heterojunction is the difference in the effective potential barrier for electrons and for holes. In Figure 2.29, for example, electrons must overcome an overall potential barrier height of $q(E_{c1} - E_{c2})$, which is smaller than the potential barrier for holes of $q(E_{v1} - E_{v2})$. This favours the injection of electrons rather than holes across the junction. Since the injection of carriers is exponentially dependent on the potential barrier height a large difference in the relative hole and electron injection rates may be obtained. The quantitative calculation of the relevant rates requires more detailed knowledge of the shape of the barrier, the effective masses of the carriers in both semiconductor materials and the possibility of tunnelling effects when a 'spike' exists such as shown in the conduction band of Figure 2.29. The Richardson–Dushman equation (Equation 2.33) is also often relevant to calculate currents flowing across these abrupt energy barriers.

Of general interest in heterojunctions is the formation of a potential well near the junction. An example of such a potential well is shown in Figure 2.29. Electrons may be trapped in this well, resulting in a thin 'sheet' of electrons, which has applications in high-frequency transistors and will therefore not be discussed in detail in this book.

In solar cells and LEDs, however, heterojunctions have important optical and electrical properties that enable high-efficiency device designs. Unwanted photon absorption can be greatly reduced by arranging the wide-gap semiconductor region to be in the path of photons entering or leaving a diode device. Regarding Figure 2.29 as an example of an LED structure, photons emitted by the recombination of electrons and holes across E_{g1} will not be significantly reabsorbed by the semiconductor of energy gap E_{g2}. This effectively reduces light loss from the LED when the light passes through the n-type semiconductor before leaving the device.

In addition, absorption or recombination may be strongly localized by engineering a region of smaller bandgap where absorption or recombination is desired. In a solar cell application, if light of a certain wavelength range reaches a junction made in narrow-gap semiconductor material by passing through a wide-gap semiconductor region then unwanted absorption of light far away from the junction can be avoided. In an LED recombination can be confined to a specific layer, which enhances recombination efficiency. These concepts

will be further described and developed in the context of solar cells and LEDs in chapters 4 and 5 respectively.

2.11   Alternating Current (AC) and Transient Behaviour

For solar cells and LED lamp applications, where steady or only slowly varying operating conditions are typical, alternating current (AC) and transient behaviour are not important. There are, however, many situations where diodes are used for switching applications and voltages and currents change rapidly with time.

There are two important operating conditions we need to discuss. Firstly in reverse bias, the diode carries only a small current and it may often be modelled as a capacitor. We shall calculate the capacitance of a p-n junction in reverse bias. It is generally sufficient to model the diode as a capacitor in circuits provided the diode remains in reverse bias; however, the value of its capacitance *changes* with reverse bias as we will show.

Secondly in forward bias, the diode builds up significant amounts of minority carrier charge on one or both sides of the depletion region. The magnitude of the minority carrier charge depends on the forward diode current; changes in forward current lead to changes in minority carrier concentrations. Since minority carriers are not generated instantly and do not recombine instantly, there is a time delay associated with the diode reaching steady state as forward bias conditions are changed. If the diode is switched between forward bias conditions and reverse bias conditions there is also a time delay involved in eliminating the minority carrier charge.

Let us consider the case of reverse bias. Capacitance C is defined by the change in charge Q on either side of an insulator caused by a given change in voltage V across the insulator or

$$C = \left| \frac{dQ}{dV} \right| \tag{2.35}$$

In a reverse-biased p-n junction diode, the insulator is effectively the depletion region of width W. The charge Q on either side of the depletion region is determined by the ionized donor or acceptor concentration. Hence from Equations 2.9 and 2.15,

$$Q = qA \frac{N_d N_a}{N_d + N_a} W$$

Using Equation 2.23 we obtain

$$Q = qA \frac{N_d N_a}{N_d + N_a} \sqrt{\frac{2\epsilon_0 \epsilon_r (V_0 - V)}{q} \left(\frac{1}{N_a} + \frac{1}{N_d} \right)}$$

$$= A \sqrt{\frac{2q\epsilon_0 \epsilon_r (V_0 - V)(N_d N_a)}{N_d + N_a}}$$

Now using Equation 2.35, and because V_0 is a constant, we have

$$C = \left| \frac{dQ}{dV} \right| = \left| \frac{dQ}{d(V_0 - V)} \right| = \frac{A}{2} \sqrt{\frac{2q\epsilon_0 \epsilon_r}{(V_0 - V)} \frac{N_d N_a}{N_d + N_a}} \tag{2.36}$$

It is clear that C is a function of the reverse bias V. This specific characteristic is taken advantage of in the *varactor diode*, which provides a variable capacitance in some circuits such as tuning circuits. Care must be taken to ensure that the diode remains in reverse bias for this application.

It is interesting to note the similarity between Equation 2.36 and the expression for capacitance of the simple parallel plate capacitor with plate separation d given by

$$C = \frac{\epsilon_0 \epsilon_r A}{d} \tag{2.37}$$

If d in Equation 2.37 is replaced with the expression for W from Equation 2.23 we obtain the identical result to Equation 2.36. This can be understood because in Equation 2.36 a small change in applied voltage dV gives rise to a small change in charge dQ near the edges of the depletion region as the depletion region width changes slightly. This is equivalent to a small change in the charge on the plates on a parallel plate capacitor due to a small change in voltage dV across the plates. See problem 2.19.

If forward bias is involved we must refer back to the excess minority carrier concentrations shown in Figure 2.16. The maximum speed with which a diode can transition between two forward bias conditions is controlled by how quickly the stored minority charge can be changed. Consider a *reduction* in forward current as a function of time. A reduction in the minority carrier concentrations as a function of time will be required (see Section 2.5).

There are two mechanisms by which the stored charge can be removed. One is by *recombination*, and the rate at which this can occur is controlled by the recombination times τ_n and τ_p. The second mechanism is by charge *withdrawal*, in which charge flows away from either side of the depletion region by diffusion into the depletion region. This second mechanism is particularly relevant if a diode is rapidly switched from forward bias to reverse bias. The electric field across the depletion region will grow and will assist these minority carriers that diffuse into the depletion region to drift across to the other side.

The simple exponential shape of the excess minority carriers shown in Figure 2.16 will be made more complex during changes in bias because minority carriers close to the depletion region will have a higher chance of withdrawal than carriers further from the depletion region edges. In contrast with this the rate of recombination is independent of position and is determined by the recombination time.

A very useful quantity that can be used to characterize this process is the *storage delay time* t_{sd}, which is the time needed for the minority carrier concentrations to reach zero after a forward-biased diode is reverse-biased. The storage delay time depends on the initial forward current flowing through the diode, and it can be measured easily since the diode voltage will reach zero when the minority carrier concentrations reach zero. Complex modelling of minority carrier recombination and withdrawal is required to accurately predict the switching time of a diode, and these models are beyond the scope of this book.

Since recombination time is an important parameter that controls switching speed in diodes, reductions in recombination times are frequently desired and may be achieved by adding traps intentionally (see Section 1.19). A common deep trap applicable to silicon diodes is gold doping added to the silicon in small concentrations in the range of 10^{14} cm^{-3} to 10^{15} cm^{-3}. This can reduce recombination times to several nanoseconds from a few microseconds without gold addition. However, adding gold does compromise diode behaviour due to unwanted carrier generation (see Section 2.8).

2.12   Summary

2.1. In a p-n junction an energy barrier is formed, which is characterized by a built-in potential V_0. The Fermi energy is constant across the junction in equilibrium. An electric field is present in the transition region of the junction.

2.2. Four currents flow in a p-n junction. In equilibrium all these currents add to zero. In forward bias the diffusion currents due to majority carriers dominate and high currents can flow. In reverse bias drift currents due to minority carriers flow but these reverse bias currents are limited in magnitude due to the limited number of minority carriers and constitute the reverse saturation current I_0.

2.3. The contact potential V_0 may be calculated from carrier concentrations and the resulting position of the Fermi energy relative to the conduction and valence band edges on either side of the junction. In equilibrium the Fermi energy is constant.

2.4. The transition region can be modelled using the depletion approximation in which a fully depleted space charge layer of width W is assumed. Values for equilibrium charge density $\rho(x)$, electric field $\varepsilon(x)$ and potential $V(x)$ result from the depletion approximation.

2.5. The diode equation derives the net diode current I obtained due to an applied potential V. It may be derived by considering the change in carrier concentrations at the edges of the depletion region implied by a change in net contact potential, and then using the diffusion equation to determine the resulting diffusion currents. Minority carrier concentrations decay exponentially with distance from the depletion region in both n-type and p-type material.

2.6. Reverse breakdown in a diode can arise from carrier avalanching. Field ionization and impact ionization in high electric fields can occur at or near the junction if the reverse bias is large enough. In addition highly doped p-n junctions may also exhibit electron tunnelling. Both mechanisms occur in the Zener diode.

2.7. If even higher doping levels are present the Fermi energy can enter the conduction and valence bands and the condition of degenerate doping is established. Tunnelling of electrons in both directions across the junction results, and a tunnelling junction results. A tunnelling junction allows efficient current flow across a p-n junction, which is important for multi-junction solar cells.

2.8. Carriers crossing over the depletion region may recombine due to deep traps. This will modify the diode equation resulting in a diode ideality factor n with values between 1 and 2. This trapping is most important at low diode currents.

2.9. Diodes require ohmic contacts to allow current to flow between metal contacts and the semiconductor. A metal-semiconductor junction may form a Schottky diode, which can be understood by thermionic emission that occurs at a metal surface. The mechanism for thermionic emission is derived to obtain the Richardson–Dushman equation. By using high doping levels tunnelling behaviour can be obtained at a metal-semiconductor junction to create a highly conductive or ohmic contact rather than a rectifying contact. A detailed analysis of metal- semiconductor junctions is influenced by surface defects, traps, dangling bonds and impurities at the metal-semiconductor interface.

2.10. Heterojunctions are formed between two semiconductors having different compositions. This permits the bandgap to change through a device. Heterojunctions are

important for both solar cells and LEDs since they have desirable electronic and optical properties for these devices such as the localization of optical absorption and the localization of electron-hole pair recombination.

2.11. An understanding of AC and transient diode behaviour requires that two mechanisms of charge storage are discussed. In reverse bias, charge is stored in the depletion region in the form of ionized donors and acceptors, which leads to diode capacitance. In forward bias, charges are stored as minority carriers on either side of the depletion region. This leads to switching delays, which can be characterized by a storage delay time. Modelling this is complex.

Suggestions for Further Reading

Neamen DA. *Semiconductor Physics and Devices*, 3rd edn. McGraw Hill, 2003.

Roulston DJ. *An Introduction to the Physics of Semiconductor Devices*. Oxford University Press, 1999.

Streetman BG and Banerjee SK. *Solid State Electronic Devices*, 6th edn. Prentice Hall, 2006.

Problems

2.1 Draw a diagram showing the excess minority carrier concentrations as a function of position in a p-n junction diode under the following conditions. Assume approximately equal doping levels on both sides of the junction:
(a) Reverse bias
(b) No bias
(c) Weak forward bias
(d) Strong forward bias
What assumption is no longer valid if forward bias current is increased beyond a certain level? Explain carefully.

2.2 An abrupt Si p-n junction has the following properties:
p-side: $N_a = 10^{18}$ cm^{-3}
n-side: $N_d = 10^{15}$ cm^{-3}
junction area $A = 10^{-4}$ cm^2
(a) Sketch a band diagram of the diode under forward bias showing hole and electron quasi-Fermi levels.
(b) Calculate the depletion region width with a reverse voltage of 10 volts.
(c) Find the diode current with the reverse bias of 10 volts. What is this current called?
(d) Find the peak value of depletion region electric field at a reverse bias of 10 volts.
(e) If the silicon exhibits avalanche breakdown at an electric field of 1×10^5 V cm^{-1} find the reverse bias voltage at which breakdown will occur at the junction.

2.3 An abrupt Si p-n junction has the following parameters:
n-side: $N_d = 5 \times 10^{18}$ cm^{-3}
p-side: $N_a = 10^{17}$ cm^{-3}
junction area $A = 10^{-2}$ cm^2
(a) Find the built-in potential V_0.
(b) Find the reverse saturation current I_0.
(c) Find the ratio of hole injection current to electron injection current at a forward current of 1 mA.
(d) Find the total minority carrier charge on each side of the diode at a forward current of 1 mA.
(e) In the p-side of the diode, at a certain distance away from the depletion region, the hole and electron currents are equal in magnitude, but opposite in direction. Find this distance.
(f) Find the quasi-Fermi level separations $F_n - F_p$ at a distance of 0.1 μm from the edges of the depletion region on either side of the junction (i.e. at a depth of 0.1 μm into the neutral p-type and n-type regions) at a forward bias of 1 mA.

2.4 An abrupt GaAs p-n junction has the following parameters:
n-side: $N_d = 2 \times 10^{18}$ cm^{-3}
p-side: $N_d = 2 \times 10^{17}$ cm^{-3}
junction area $A = 10^{-4}$ cm^2
(a) Find the built-in potential V_0.
(b) Find the ratio of hole current to electron current crossing the junction at a total forward current of 5 mA.
(c) Find the reverse saturation current I_0.

(d) At a forward current of 5 mA, find the total minority carrier charge on each side of the diode.

(e) Find the distance into the p-side at which the minority electron concentration is half the maximum value. Sketch it as a function of distance into the p-side starting from the edge of the depletion region.

2.5 Calculate the electric field as a function of position in a p-n junction having a depletion region of width W and constant doping levels of N_d and N_a on the n- and p-sides respectively. To do this, first find the width of the depletion region on either side of the junction as a function of W. Then use Gauss's law to determine the electric field at any point in the depletion region by considering a Gaussian surface that covers only a fraction of the space charge on either side of the junction. Show that the electric field increases linearly and reaches a maximum value at the junction when the Gaussian surface encloses all the space charge on one side of the junction. Sketch the field as a function of position.

2.6 If the diode of Problem 2.4 is forward biased, sketch how the electric field would vary as a function of position throughout the depletion region. Repeat for a reverse bias. Compare these sketches to the sketch without bias.

2.7 A GaAs diode is reverse biased and it exhibits an increasing reverse current as bias increases. Explain how this occurs based on generation or recombination in the depletion region. What changes would be made to the properties of the semiconductor to reduce this effect?

2.8 A tunnel diode is formed using highly doped silicon with a total depletion width under equilibrium conditions of 5 nm. According to the depletion approximation, what doping level would be needed to achieve this? Assume equal doping levels on both sides of the junction.

2.9 The forward current in a silicon diode increases with bias voltage with the following relationship:

$$I \propto \exp\left(\frac{qV}{1.5kT}\right)$$

(a) Explain the physics underlying this dependence. Plot a representative graph of the current versus voltage dependence using linear x- and y-axes and compare its shape to the graph of a diode having a relationship that obeys the diode equation.

(b) At low forward bias voltages a diode is observed to behave according to $I \propto \exp(\frac{qV}{1.5kT})$ but at higher voltages its current–voltage dependence approaches the diode equation. Explain carefully.

(c) A silicon diode at a given forward bias voltage range behaves according to $I \propto \exp(\frac{qV}{2kT})$ at a junction temperature of $100°C$ but at $-50°C$ its current–voltage dependence for the same bias voltage range follows the diode equation. Explain carefully.

2.10 Figure 2.27 shows the band structure of an effective ohmic contact that works by tunnelling applied to n-type silicon.

(a) Sketch the analogous band structure for an ohmic contact applied to p-type silicon that relies on tunnelling.

(b) Propose a metal contact material that forms a tunnelling type ohmic contact to n-type GaAs. Look in the literature and see if your answer is a good prediction of materials used in practice.

(c) Repeat (b) for p-type GaAs.

2.11 A Schottky diode is composed of a junction between p-type silicon and aluminium. The barrier height E_b is 0.4 eV and junction area is 1000 μm^2.

(a) Calculate I_e at room temperature.

(b) Find and plot the diode current as a function of applied voltage at room temperature for both forward and reverse bias. You may neglect reverse breakdown.

2.12 (a) Suggest why the boundary conditions for travelling waves relevant to the Schottky diode applied to Equation 1.27 ($k_x = \frac{2n_x\pi}{a}$ etc.) are a factor of two different from the boundary condition for standing waves. Using the travelling wave boundary conditions, show that you obtain the number of available electron states per unit volume of crystal as $dn = \frac{2}{h^2}dp_x dp_y dp_z$.

(b) *The Born–von Karman boundary condition* is relevant here and is the basis for the travelling wave boundary condition. Look up this boundary condition and explain it in the context of the travelling wave boundary condition.

2.13 A silicon p-n junction diode has the following parameters: $N_d = 2 \times 10^{18}$ cm^{-3}, $N_a = 2 \times 10^{16}$ cm^{-3}, $\tau_n = \tau_p = 2 \times 10^{-6}$ s, $D_n = 25$ cm^2 s^{-1} and $D_p = 8$ cm^2 s^{-1}. A light source is incident only on the depletion region, producing a generation current density of $J_{gen} = 50$ mA cm^{-2}. The diode is open circuited. The generation current density forward biases the junction, inducing a forward-bias current in the opposite direction to the generation current. A steady-state condition is reached when the generation current density and forward-bias current density are equal in magnitude. What is the induced forward-bias voltage at this steady-state condition?

2.14 An abrupt Si p-n junction has the following parameters:

p-side: $N_a = 10^{17}$ cm^{-3}

n-side: $N_d = 10^{14}$ cm^{-3}

junction area $A = 1 \times 10^{-5}$ cm^2

  Find:

(a) V_0, the built-in potential.

(b) I_0, the reverse saturation current.

(c) The depletion region width at 0 volts.

(d) The depletion region width at 10 volts reverse bias.

(e) The peak electric field at 10 volts reverse bias.

(f) The ratio of hole-to-electron current flow in forward bias.

(g) The diode capacitance at reverse bias of 5 volts, 10 volts and 15 volts.

2.15 An abrupt Si p-n junction has the following properties:

p-side: $N_a = 10^{17}$ cm^{-3},

n-side: $N_d = 10^{15}$ cm^{-3}

junction area $A = 10^{-4}$ cm^2

(a) Sketch an equilibrium band diagram showing E_f and V_0.

(b) Calculate V_0.

(c) Calculate the space charge width with zero applied voltage.

(d) Find the maximum electric field at a reverse bias of 10 volts.

(e) Find I_o, the diode reverse saturation current.

   (f) Find the breakdown field for silicon if the diode has a reverse breakdown voltage
       of 100 volts. Hint: Use the highest field in the depletion region for this calculation.
   (g) Find the depletion region width just before reverse breakdown.
   (h) Find the diode capacitance at a reverse bias of 10 volts.
   (i) Find the voltage across the diode at a forward current at 1 A.

2.16 An abrupt Si p-n junction has the following properties:

   p-side: $N_a = 10^{18}$ cm^{-3}
   n-side: $N_d = 10^{16}$ cm^{-3}
   junction area $A = 10^{-4}$ cm^2

   (a) Sketch a band diagram of a diode under forward bias showing hole and electron
       quasi-Fermi levels.
   (b) Calculate the space charge width with a reverse voltage of 10 volts.
   (c) Find the diode capacitance at an applied voltage of 10 volts reverse bias.
   (d) Find the diode current with a reverse bias of 10 volts.
   (e) Find the ratio of hole to electron current that crosses over the depletion region.
   (f) Find the peak value of depletion region electric field at a reverse bias of 10 volts.

2.17 In a p$^+$-n junction at room temperature, the n-doping N_d is doubled. How do the
   following two parameters change if everything else is unchanged?
   (a) breakdown voltage
   (b) built-in voltage

2.18 Ohmic contacts are needed on a p-n junction made with silicon. Metal contact pads
   will be deposited on the p and n regions, and then diffused in for a short time to form
   ohmic contacts. List some suitable materials for p and n ohmic contacts.

2.19 Show that you can obtain expression 2.36 from 2.37. Sketch the depletion region of a
   diode and show the locations of differential charge dQ obtained by a small change in
   applied voltage dV.

3

Photon Emission and Absorption

Objectives

1. Introduce the basic forms of luminescence.
2. Present the dipole model of luminescence based on radiation from the acceleration of charges.
3. Introduce the quantum mechanical description of acceleration of charges that can be used to calculate radiation rate and radiative power.
4. Introduce the free exciton and the mechanism by which an exciton emits a photon through dipole radiation.
5. Describe the two-electron atom and the concept of indistinguishable particles.
6. Present the resulting molecular excitons and their classification as singlet excitons or triplet excitons.

7. Describe the luminescent properties of fluorescence and phosphorescence that are observed from singlet and triplet excitons respectively.
8. Describe the band-to-band emission and recombination model that determines absorption or radiation spectra based on band states and their probability of occupancy.
9. Introduce the human visual system and the units of luminescence and colour that allow light sources to be described with relevance to human perception.

3.1 Introduction to Luminescence and Absorption

We have discussed the electronic aspects of a p-n junction in some detail. However, the processes by which light is absorbed or emitted are a crucial aspect of solar cells and LEDs. The p-n junction is an ideal device to either absorb or emit photons, and these processes occur when an electron-hole pair is generated or annihilated respectively. The p-n diode efficiently transports both electrons and holes either towards the junction for photon generation to occur, or away from the junction for electric current generation due to photon absorption to occur.

We will discuss in some detail the theory of *luminescence*, in which a photon is created by an electron-hole pair. These concepts can be readily understood in reverse to explain photon absorption also.

Technologically important forms of luminescence may be broken down into several categories as shown in Table 3.1. Although the means by which the luminescence is excited varies, *all luminescence is generated by means of accelerating charges*. The portion of the electromagnetic spectrum visible to the human eye contains wavelengths from 400 to 700 nm. The evolution of the relatively narrow sensitivity range of the human eye is a

Table 3.1 *Luminescence types, applications and typical efficiencies. Efficiency (η) is given in visible light output power as a fraction of input power*

Luminescence type	Examples
Blackbody radiation (light generated due to the temperature of a body)	Sun Tungsten filament lamp ($\eta = 5\%$)
Photoluminescence (light emitted by a material that is stimulated by electromagnetic radiation)	Fluorescent lamp phosphors ($\eta = 80\%$)
Cathodoluminescence (light emitted by a material that is excited by high-energy electrons incident from a vacuum)	Cathode-ray tube screen ($\eta = 25\%$)
Electroluminescence (light emitted by a material that is directly electrically excited)	Light emitting diode ($\eta = 50\%$)

complex subject, but is intimately related to the solar spectrum, the absorbing behaviour of the terrestrial atmosphere, and the reflecting properties of terrestrial materials, green lying near the middle of the useful spectrum. Not surprisingly, the wavelength at which the human eye is most sensitive is also green.

Visible light emission is the most important wavelength range for both organic and inorganic LEDs since LEDs are heavily used for lighting and display applications. The display applications typically include red, green and blue wavelengths in a *trichromatic* scheme that allows humans to perceive a wide range of colours from a set of only three primary colours.

Infrared (IR) and ultraviolet (UV) radiation must also be considered for both solar cells and LEDs that are not only intended for visible light emission or absorption. The sun includes these IR and UV wavelengths, and solar cells in particular are heavily dependent on IR absorption. Infrared LEDs are well developed and are used for remote control and sensing applications. More recently UV LEDs are being introduced for industrial processing applications.

3.2   Physics of Light Emission

In order to understand the processes of light emission and light absorption in more detail we will examine the behaviour of charges and moving charges. A stationary point charge q results in electric field lines that emanate from the charge in a radial geometry as shown in Figure 3.1. A charge moving with a uniform velocity relative to an observer gives rise to a magnetic field. Figure 3.2 shows the resulting magnetic field when the point charge moves away from an observer.

Since both electric and magnetic fields store energy the total energy density is given by

$$E = \frac{1}{2\epsilon_0}\varepsilon^2 + \frac{1}{2\mu_0}B^2$$

It is important to note that this energy field falls off in density as we move away from the charge but it moves along with the charge provided that the charge is either stationary or undergoing uniform motion. There is no flow of energy from the charge.

For a charge undergoing *acceleration*, however, energy continuously travels away from the charge. Consider the charge q in Figure 3.3. Assume it is initially at rest in position A, then accelerates to position B, and stops there. The electric field lines now emanate from
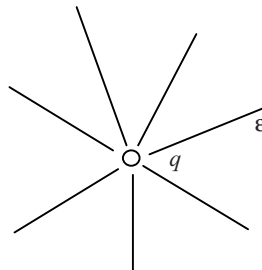
Figure 3.1   *Lines of electric field ε due to a point charge q*
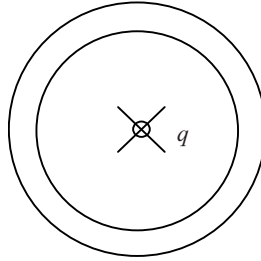
Figure 3.2 *Closed lines of magnetic field B due to a point charge q moving into the page with uniform velocity*

position B, but further out the lines had emanated from position A. The field lines cannot convey information about the location of the charge at speeds greater than the velocity of light c. This results in kinks in the lines of electric field, which propagate away from q with velocity c. Each time q accelerates, a new series of propagating kinks is generated. Each kink is made up of a component of ε that is transverse to the direction of expansion, which we call $\varepsilon_\perp$. If the velocity of the charge during its acceleration does not exceed a small fraction of c, then for large distances away from charge q,

$$\varepsilon_\perp = \frac{qa}{4\pi\epsilon_0 c^2 r}\sin\theta$$

Here, a is acceleration, r is the radial distance between the charge and the position where the electric field is evaluated, and θ is the angle between the direction of acceleration and the radial direction of the transverse field. The strongest transverse field occurs in directions normal to the direction of acceleration, as shown by Figure 3.3.
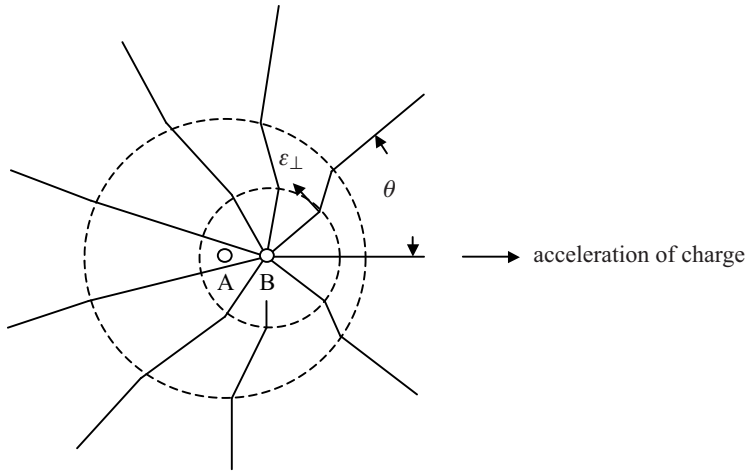
Figure 3.3 *Lines of electric field emanating from an accelerating charge*
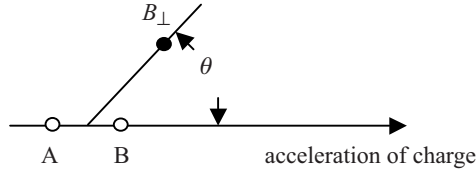
Figure 3.4 *Direction of magnetic field $B_\perp$ emanating from an accelerating charge. $B_\perp$ is perpendicular to both acceleration and the radial direction*

Likewise, a transverse magnetic field $B_\perp$ that points in a direction perpendicular to both acceleration and the radial direction is generated during the acceleration of the charge, as shown in Figure 3.4, and is given by

$$B_\perp = \frac{\mu_0 q a}{4\pi c r} \sin\theta$$

The two transverse fields propagate outwards with velocity c each time q undergoes acceleration, giving rise to the electromagnetic radiation. Note that $\varepsilon_\perp$ and $B_\perp$ are perpendicular to each other. The energy density of the radiation is

$$E = \frac{1}{2\epsilon_0}\varepsilon_\perp^2 + \frac{1}{2\mu_0}B_\perp^2$$

The Poynting vector, or energy flow per unit area (radiation intensity), is

$$\vec{S} = \frac{1}{\mu_0}\vec{\varepsilon}_\perp \times \vec{B}_\perp$$

$$= \frac{q^2 a^2}{16\pi^2\epsilon_0 c^3 r^2}\sin^2\theta\,\hat{r}$$

where $\hat{r}$ is a unit radial vector.

Maximum energy is emitted in a ring perpendicular to the direction of acceleration, and none is emitted along the direction of acceleration. To obtain the *total* radiated energy per unit time or power P leaving q due to its acceleration, we integrate S over a sphere surrounding q to obtain

$$P = \int_{\text{sphere}} S(\theta)\,\mathrm{d}A = \int_0^{2\pi}\int_0^\pi S(\theta)\,r^2\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi$$

or

$$P = \int_0^\pi S(\theta)2\pi r^2 \sin\theta\,\mathrm{d}\theta$$

Substituting for $S(\theta)$ we obtain

$$P = \frac{1}{16\pi\epsilon_0}\frac{2q^2 a^2}{c^3}\int_0^\pi \sin^3\theta\,\mathrm{d}\theta$$

which can be integrated (see Problem 3.19) to obtain

$$P = \frac{1}{4\pi\epsilon_0} \frac{2q^2a^2}{3c^3}$$

(3.1)

3.3   Simple Harmonic Radiator

If a charge q oscillates about the origin along the x-axis and its position is given by $x = A \sin \omega t$ then we can calculate the average power radiated away from the oscillating charge. Note that the acceleration a of the charge is given by

$$a = \frac{d^2x}{dt^2} = -A\omega^2 \sin \omega t$$

and using Equation 3.1

$$P = \frac{2q^2 A^2 \omega^4 \sin^2 \omega t}{12\pi\epsilon_0 c^3}$$

which varies with time as $\sin^2 \omega t$.

To obtain *average* power we integrate over one cycle to obtain

$$\overline{P} = \frac{\omega}{2\pi} \frac{2q^2 A^2 \omega^4}{12\pi\epsilon_0 c^3} \int_0^{\frac{2\pi}{\omega}} \sin^2 \omega t \ dt$$

which yields

$$\overline{P} = \frac{q^2 A^2 \omega^4}{12\pi\epsilon_0 c^3}$$

(3.2)

If we now consider that an equal and opposite stationary charge $-q$ is located at $x = 0$ then we have a dipole radiator with electric dipole moment of amplitude $p = qA$. Now we may write

$$\overline{P} = \frac{p^2 \omega^4}{12\pi\epsilon_0 c^3}$$

(3.3)

Radiation that does not rely on dipoles also exists. For example, a synchrotron radiation source is an example of a radiator that relies on the constant centripetal acceleration of an orbiting charge. In a synchrotron the acceleration is in the direction of the orbit radius pointing to the centre of the orbit and the radiation is strongest in a direction tangential to the orbit. There are also quadrupoles, magnetic dipoles and other oscillating charge configurations that do not comprise dipoles; however, they have much lower rates of energy release and by far the dominant form of radiation is from dipoles.

An electron and a hole effectively behave like a dipole radiator when they recombine to create a photon. The photon is not created instantly, and many oscillations of the charges must occur before the photon is fully formed. The description of the photon as a wave packet is very relevant since the required photon energy is gradually built up as the dipole oscillates to complete the wave packet. Unless the wave packet is fully formed no photon exists; the smallest unit of electromagnetic radiation is the photon.

Since we need to describe the positional behaviour of electrons and holes using quantum concepts we will now proceed to introduce a simple quantum expression to incorporate in the calculation of dipole radiation.

3.4   Quantum Description

A charge q does not exhibit energy loss or radiation when in a *stationary* state or eigenstate of a potential energy field. This means, in a stationary state, that no net acceleration of the charge occurs, in spite of its uncertainty in position and momentum dictated by the Heisenberg uncertainty principle. Experience tells us, however, that radiation may be produced when a charge moves from one stationary state to another; it will be the purpose of this section to show that *radiation is produced if an oscillating dipole results from a charge moving from one stationary state to another.*

Consider a charge q initially in normalized stationary state ψ_n and eventually in normalized stationary state $\psi_{n'}$. During the transition, a superposition state is created which we shall call ψ_s:

$$\psi_s = a\psi_n + b\psi_{n'}$$

where

$$|a|^2 + |b|^2 = 1$$

to normalize the superposition state. Here a and b are time-dependent coefficients. Initially $a = 1$ and $b = 0$ and finally $a = 0$ and $b = 1$.

Quantum mechanics allows us to calculate the time-dependent expected value of the position $\langle r \rangle (t)$ of a particle in a quantum state. For example, for stationary state ψ_n

$$\langle r \rangle (t) = \langle \psi_n |r| \psi_n \rangle = \int_V |\psi_n|^2 \, r \, dV$$

where V represents all space. For any stationary state

$$\psi (r, \, t) = \phi(r) \exp \frac{iE}{\hbar} t$$

and hence

$$\langle r \rangle (t) = \int_V \left[\phi(r) \exp \frac{iE}{\hbar} t \phi(r) \exp \frac{-iE}{\hbar} t \right] r \, dV = \int_V r \phi^2(r) \, dV$$

This expression for $\langle r \rangle$ is therefore not a function of time, which is the fundamental idea underlying the name *stationary* state. A stationary state does not radiate and there is no energy loss associated with the behaviour of an electron in such a state. Note that electrons are not truly stationary in a quantum state. It is therefore the quantum state that is described as stationary and not the electron itself. Quantum mechanics sanctions the existence of a charge that can move around in a small space but that does not possess a measurable acceleration. Classical physics fails to describe or predict this.

If we now calculate the expectation value of the position of q for the superposition state ψ_s in the same manner we obtain

$$\langle r \rangle_s = \langle a\psi_n + b\psi_{n'} \,|\, r \,|\, a\psi_n + b\psi_{n'} \rangle$$
$$= |a|^2 \langle \psi_n \,|r|\, \psi_n \rangle + |b|^2 \langle \psi_{n'} \,|r|\, \psi_{n'} \rangle + a^*b \langle \psi_n \,|r|\, \psi_{n'} \rangle + b^*a \langle \psi_{n'} \,|r|\, \psi_n \rangle$$

As before the time-dependent component is written separately from the spatial component:

$$\psi_n = \phi_n \exp\left(-i\frac{E_n}{\hbar}t\right)$$

where ϕ_n is the spatially dependent part of ψ_n. Hence, disregarding the time-independent terms, which do not contribute to radiation, we obtain

$$\langle r \rangle_s(t) = a^*b \langle \phi_n \,|r|\, \phi_{n'} \rangle \exp\left[\frac{i(E_n - E_{n'})t}{\hbar}\right] + b^*a \langle \phi_{n'} \,|r|\, \phi_n \rangle \exp\left[\frac{i(E_n - E_{n'})t}{\hbar}\right]$$
$$= 2\,\mathrm{Re}\left\{a^*b \langle \phi_n \,|r|\, \phi_{n'} \rangle \exp\left[\frac{i(E_n - E_{n'})t}{\hbar}\right]\right\}$$

since the position must be a real number. This may be written using Euler's formula as

$$\langle r \rangle_s(t) = 2\left|a^*b \langle \phi_n \,|r|\, \phi_{n'} \rangle\right| \cos(\omega_{nn'}t + \delta)$$
$$= 2\,|r_{nn'}| \cos(\omega_{nn'}t + \delta) \qquad (3.4)$$

Note that we have introduced the relationship $E = \hbar\omega$ that defines the energy of one photon generated by the charge q as it moves from ψ_n to $\psi_{n'}$. Note also that $\langle r \rangle_s(t)$ is oscillating with frequency $\omega_{nn'} = \dfrac{E_n - E_{n'}}{\hbar}$ such that the required number of oscillations at the required frequency releases one photon having energy $E = \hbar\omega_{nn'}$ from the oscillating charge. The term $|r_{nn'}|$ also varies with time, but does so slowly compared with the cosine term. This is illustrated in Figure 3.5.
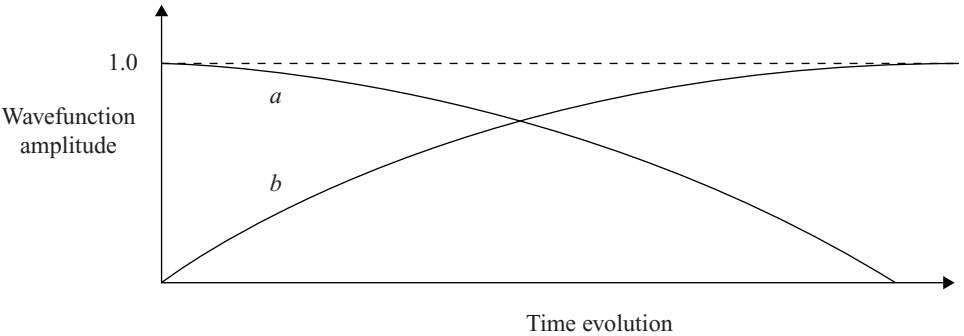
Figure 3.5   *A time-dependent plot of coefficients* a *and* b *is consistent with the time evolution of wavefunctions* ϕ_n *and* $\phi_{n'}$. *At t = 0, a = 1 and b = 0. Next a superposition state is formed during the transition such that* $|a|^2 + |b|^2 = 1$. *Finally after the transition is complete a = 0 and b = 1*

Example 3.1

An electron within a dipole oscillates about $x = 0$ with amplitude $A = 1$ Å to produce a photon with $\lambda = 550$ nm.

(a) Find the radiation power in watts.
(b) Find the photon energy.
(c) Find the approximate time taken to release one photon.
(d) Find the approximate number of oscillations of the electron required to produce one photon or wave packet.

Solution

(a) From the classical wave equation

$$c = f\lambda = \frac{\omega}{2\pi}\lambda$$

we can solve for ω to obtain

$$\omega = \frac{2\pi c}{\lambda} = \frac{2\pi \times 3.0 \times 10^8 \text{ m s}^{-1}}{550 \times 10^{-9} \text{ m}}$$

now,

$$\bar{P} = \frac{q^2 A^2 \omega^4}{12\pi \epsilon_0 c^3} = \frac{(1.6 \times 10^{-19}\text{ C})^2 \times (10^{-10} \text{ m})^2 \times (2\pi)^4 \times (3 \times 10^8 \text{ m s}^{-1})}{12\pi (8.85 \times 10^{-12} \text{ F m}^{-1})(550 \times 10^{-9} \text{ m})^4}$$

$$= 4 \times 10^{-12} \text{ W}$$

(b) One photon of this wavelength has energy

$$E_{\text{photon}} = \frac{hc}{\lambda} = \frac{6.62 \times 10^{-34} \text{ Js} \times 3 \times 10^8 \text{ m s}^{-1}}{550 \times 10^{-9} \text{ m}} = 3 \times 10^{-19} \text{ J}$$

(c) Hence, the approximate length of time taken to release the photon is

$$T = \frac{E_{\text{photon}}}{\bar{P}} = \frac{3 \times 10^{-19} \text{ J}}{4 \times 10^{-12} \text{ W}} = 7.7 \times 10^{-8} \text{ s}$$

Note that this result assumed that the rate of radiation is steady.

(d) Since the period of electromagnetic oscillation is

$$T_{\text{oscillation}} = \frac{\lambda}{c} = \frac{550 \times 10^{-9} \text{ m}}{3 \times 10^8 \text{ m s}^{-1}} = 1.8 \times 10^{-15} \text{ s}$$

the number of oscillations that take place during the time required to release the photon becomes

$$N = \frac{T}{T_{\text{oscillation}}} = \frac{7.7 \times 10^{-8} \text{ s}}{1.8 \times 10^{-15} \text{ s}} = 4.3 \times 10^7 \text{ oscillations}$$

We can regard these oscillations as being incorporated into the photon wave packet. We have assumed $|r_{nn'}|$ to be a constant.

We may also define a photon emission rate $R_{nn'}$ of a continuously oscillating charge q. We use Equations 3.2 and 3.4 and $E = \hbar\omega$ to obtain

$$R_{nn'} = \frac{\bar{P}}{\hbar\omega} = \frac{q^2\omega^3}{3\pi\epsilon_0 c^3 \hbar}|r_{nn'}|^2 \frac{\text{photons}}{\text{second}}$$

We are now particularly interested in dipoles formed from a hole-electron pair. A hole-electron pair may produce one photon before it is annihilated, which leads us to examine the hole-electron pair in more detail. Also of relevance is photon absorption, in which a hole-electron pair is created due to photon absorption.

3.5   The Exciton

A hole and an electron can exist as a valence band state and a conduction band state. In this model the two particles are not localized and they are both represented using Bloch functions in the periodic potential of the crystal lattice. If the mutual attraction between the two becomes significant then a new description is required for their quantum states that is valid before they recombine but after they experience some mutual attraction.

The hole and electron can exist in quantum states that are actually *within* the energy gap. The band model in Chapter 1 does not consider this situation. Just as a hydrogen atom consists of a series of energy levels associated with the allowed quantum states of a proton and an electron, a series of energy levels associated with the quantum states of a hole and an electron also exists. This hole-electron entity is called an *exciton*, and the exciton behaves in a manner that is similar to a hydrogen atom with one important exception: a hydrogen atom has a lowest energy state or ground state when its quantum number $n = 1$, but a exciton, which also has a ground state at $n = 1$, has an opportunity to be annihilated when the electron and hole eventually recombine.

The energy levels and Bohr radius for a hydrogen atom were presented in Section 1.13. For an exciton we need to modify the electron mass m to become the reduced mass μ of the hole-electron pair, which is given by

$$\frac{1}{\mu} = \frac{1}{m_e^*} + \frac{1}{m_h^*}$$

For direct gap semiconductors such as GaAs this is about one order of magnitude smaller than the free electron mass m. In addition the exciton exists inside a semiconductor rather than in a vacuum. The relative dielectric constant ϵ_r must be considered, and it is approximately 10 for typical inorganic semiconductors. Now using Equation 1.39 we have the ground state energy for an exciton of

$$E_{\text{exciton}} = \frac{-\mu q^4}{8\epsilon_o^2 \epsilon_r^2 h^2} \cong \frac{E_{\text{Rydberg}}}{1000}$$

This yields a typical exciton ionization energy or *binding energy* of under 0.1 eV. Using Equation 1.40 the exciton radius in the ground state ($n = 1$) will be given by

$$a_{\text{exciton}} = \frac{4\pi\epsilon_0 \, \epsilon_r \hbar^2}{\mu q^2} \cong 100 a_0$$

which yields an exciton radius of the order of 50 Å. Since this radius is much larger than the lattice constant of a semiconductor, we are justified in our use of the bulk semiconductor parameters for effective mass and relative dielectric constant.

Our picture is now of a hydrogen atom-like entity drifting around within the semiconductor crystal and having a series of energy levels analogous to those in a hydrogen atom. Just as a hydrogen atom has energy levels $E_n = \frac{13.6}{n^2}$ eV where quantum number n is an integer, the exciton has similar energy levels but in a much smaller energy range, and a quantum number $n_{exciton}$ is used.

The exciton must transfer energy to be annihilated. When an electron and a hole form an exciton it is expected that they are initially in a high energy level with a large quantum number $n_{exciton}$. This forms a larger, less tightly bound exciton. Through *thermalization* the exciton loses energy to lattice vibrations and approaches its ground state. Its radius decreases as $n_{exciton}$ approaches 1. Once the exciton is more tightly bound and $n_{exciton}$ is a small integer, the hole and electron can then form an effective dipole and radiation may be produced to account for the remaining energy and to annihilate the exciton through the process of dipole radiation. When energy is released as electromagnetic radiation, we can determine whether or not a particular transition is allowed by calculating the term $|r_{nn'}|$ from Equation 3.4 and determining whether it is zero or non-zero. If $|r_{nn'}| = 0$ then this is equivalent to saying that dipole radiation will *not* take place and a photon cannot be created. Instead lattice vibrations remove the energy. If $|r_{nn'}| > 0$ then this is equivalent to saying that dipole radiation *can* take place and a photon can be created. We can represent the exciton energy levels in a semiconductor as shown in Figure 3.6.

At low temperatures the emission and absorption wavelengths of electron-hole pairs must be understood in the context of excitons in all p-n junctions. The existence of excitons, however, is generally hidden at room temperature and at higher temperatures in *inorganic* semiconductors because of the temperature of operation of the device. The exciton is not stable enough to form from the distributed band states and at room temperature kT may be larger than the exciton energy levels. In this case the spectral features associated with excitons will be masked and direct gap or indirect gap band-to-band transitions
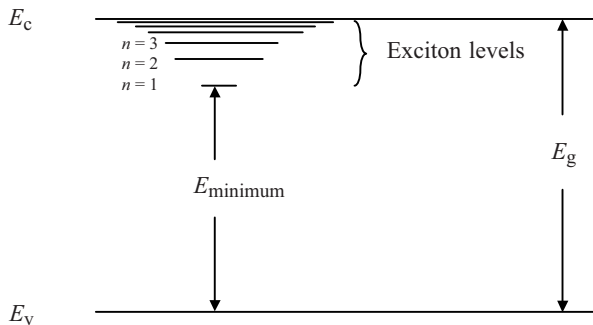
Figure 3.6 *The exciton forms a series of closely spaced hydrogen-like energy levels that extend inside the energy gap of a semiconductor. If an electron falls into the lowest energy state of the exciton corresponding to* n = 1 *then the remaining energy available for a photon is* $E_{minimum}$
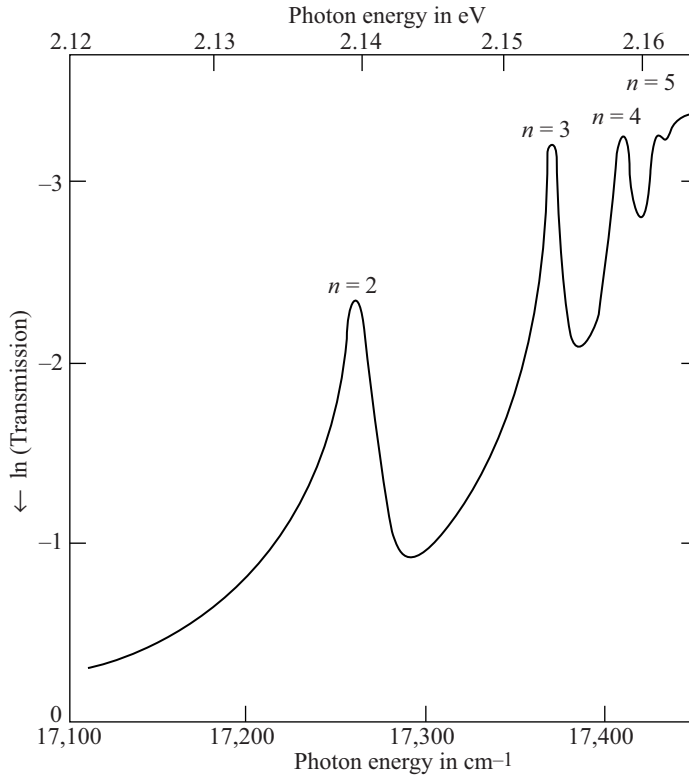
Figure 3.7   *Low-temperature transmission as a function of photon energy for Cu_2O. The absorption of photons is caused through excitons, which are excited into higher energy levels as the absorption process takes place. Cu_2O is a semiconductor with a bandgap of 2.17 eV. Reprinted from Kittel, C., Introduction to Solid State Physics, 6e, ISBN 0-471-87474-4. Copyright (1986) John Wiley and Sons, Australia*

occur. Nevertheless photoluminescence or absorption measurements at low temperatures conveniently provided in the laboratory using liquid nitrogen (77 K) or liquid helium (4.2 K) clearly show exciton features, and excitons have become an important tool to study inorganic semiconductor behaviour. An example of the transmission as a function of photon energy of a semiconductor at low temperature due to excitons is shown in Figure 3.7.

In an indirect gap inorganic semiconductor at room temperature without the formation of excitons, the hole-electron pair can lose energy to phonons and be annihilated but not through dipole radiation. In a direct-gap semiconductor, however, dipole radiation can occur. The calculation of $|r_{nn'}|$ is also relevant to band-to-band transitions. Since a dipole does not carry linear momentum it does not allow for the conservation of electron momentum during electron-hole pair recombination in an indirect gap semiconductor crystal and dipole radiation is forbidden. The requirement of a direct gap for a band-to-band transition that conserves momentum is consistent with the requirements of dipole radiation. Dipole radiation is effectively either allowed or forbidden in band-to-band transitions.

Not all excitons are free to move around in the semiconductor. *Bound excitons* are often formed that associate themselves with defects in a semiconductor crystal such as vacancies and impurities. In organic semiconductors *molecular excitons* form, which are very important for an understanding of optical processes that occur in organic semiconductors. This is because molecular excitons typically have high binding energies of approximately 0.4 eV. The reason for the higher binding energy is the confinement of the molecular exciton to smaller spatial dimensions imposed by the size of the molecule. This keeps the hole and electron closer and increases the binding energy compared to free excitons. In contrast to the situation in inorganic semiconductors, molecular excitons are thermally stable at room temperature and they generally determine emission and absorption characteristics of organic semiconductors in operation. The molecular exciton will be discussed in Section 3.7. We will first need to discuss in more detail the physics required to understand excitons and optical processes in molecular materials.

3.6   Two-Electron Atoms

Until now we have focused on dipole radiators that are composed of two charges, one positive and one negative. In Section 3.3 we introduced an oscillating dipole having one positive charge and one negative charge. In Section 3.5 we discussed the exciton, which also has one positive charge and one negative charge.

However, we also need to understand radiation from molecular systems with two or more electrons, which form the basis of organic semiconductors. Once a system has two or more identical particles (electrons) there are additional and very fundamental quantum effects that we need to consider. In inorganic semiconductors, band theory gives us the tools to handle large numbers of electrons in a periodic potential. In organic semiconductors electrons are confined to discrete organic molecules and 'hop' from molecule to molecule. Band theory is still relevant to electron behaviour within a given molecule provided it contains repeating structural units (see Chapter 6).

Nevertheless we need to study the electronic properties of molecules more carefully because molecules contain multiple electrons, and exciton properties in molecules are rather different from the excitons we have discussed in inorganic semiconductors. The best starting point is the helium atom, which has a nucleus with a charge of $+2q$ as well as two electrons each with a charge of $-q$. A straightforward solution to the helium atom using Schrödinger's equation is not possible since this is a *three-body* system; however, we can understand the behaviour of such a system by applying the Pauli exclusion principle and by including the spin states of the two electrons.

When two electrons at least partly overlap spatially with one another their wavefunctions must conform to the Pauli exclusion principle; however, there is an additional requirement that must be satisfied. The two electrons must be carefully treated as *indistinguishable* because once they have even a small spatial overlap there is no way to know which electron is which. We can only determine a probability density $|\psi|^2 = \psi^*\psi$ for each wavefunction but we cannot determine the precise location of either electron at any instant in time and therefore there is always a chance that the electrons exchange places. There is no way to label or otherwise identify each electron and the wavefunctions must therefore not be specific about the identity of each electron.

If we start with Schrödinger's equation and write it by adding up the energy terms from the two electrons we obtain

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2 \psi_T}{\partial x_1^2}+\frac{\partial^2 \psi_T}{\partial y_1^2}+\frac{\partial^2 \psi_T}{\partial z_1^2}\right)-\frac{\hbar^2}{2m}\left(\frac{\partial^2 \psi_T}{\partial x_2^2}+\frac{\partial^2 \psi_T}{\partial y_2^2}+\frac{\partial^2 \psi_T}{\partial z_2^2}\right)+V_T\psi_T=E_T\psi_T$$

(3.5)

Here $\psi_T(x_1, y_1, z_1, x_2, y_2, z_2)$ is the wavefunction of the two-electron system, $V_T(x_1, y_1, z_1, x_2, y_2, z_2)$ is the potential energy for the two-electron system and E_T is the total energy of the two-electron system. The spatial coordinates of the two electrons are (x_1, y_1, z_1) and (x_2, y_2, z_2).

To simplify our treatment of the two electrons we will start by assuming that the electrons do not interact with each other. This means that we are neglecting coulomb repulsion between the electrons. The potential energy of the total system is then simply the sum of the potential energy of each electron under the influence of the helium nucleus. Now the potential energy can be expressed as the sum of two identical potential energy functions $V(x, y, z)$ for the two electrons and we can write

$$V_T(x_1, y_1, z_1, x_2, y_2, z_2) = V(x_1, y_1, z_1) + V(x_2, y_2, z_2)$$

Substituting this into Equation 3.5 we obtain

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2 \psi_T}{\partial x_1^2}+\frac{\partial^2 \psi_T}{\partial y_1^2}+\frac{\partial^2 \psi_T}{\partial z_1^2}\right)-\frac{\hbar^2}{2m}\left(\frac{\partial^2 \psi_T}{\partial x_2^2}+\frac{\partial^2 \psi_T}{\partial y_2^2}+\frac{\partial^2 \psi_T}{\partial z_2^2}\right)$$
$$+V(x_1, y_1, z_1)\psi_T+V(x_2, y_2, z_2)\psi_T=E_T\psi_T$$

(3.6)

If we look for solutions for ψ_T of the form $\psi_T = \psi(x_1, y_1, z_1)\psi(x_2, y_2, z_2)$ then Equation 3.6 becomes

$$-\frac{\hbar^2}{2m}\psi(x_2, y_2, z_2)\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial y_1^2}+\frac{\partial^2}{\partial z_1^2}\right)\psi(x_1, y_1, z_1)$$
$$-\frac{\hbar^2}{2m}\psi(x_1, y_1, z_1)\left(\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial y_2^2}+\frac{\partial^2}{\partial z_2^2}\right)\psi(x_2, y_2, z_2)$$
$$+V(x_1, y_1, z_1)\psi(x_1, y_1, z_1)\psi(x_2, y_2, z_2)$$
$$+V(x_2, y_2, z_2)\psi(x_1, y_1, z_1)\psi(x_2, y_2, z_2)$$
$$=E_T\psi(x_1, y_1, z_1)\psi(x_2, y_2, z_2)$$

(3.7)

Dividing Equation 3.7 by $\psi(x_1, y_1, z_1)\psi(x_2, y_2, z_2)$ we obtain

$$-\frac{\hbar^2}{2m}\frac{1}{\psi(x_1, y_1, z_1)}\left(\frac{\partial^2}{\partial x_1^2}+\frac{\partial^2}{\partial y_1^2}+\frac{\partial^2}{\partial z_1^2}\right)\psi(x_1, y_1, z_1)$$
$$-\frac{\hbar^2}{2m}\frac{1}{\psi(x_2, y_2, z_2)}\left(\frac{\partial^2}{\partial x_2^2}+\frac{\partial^2}{\partial y_2^2}+\frac{\partial^2}{\partial z_2^2}\right)\psi(x_2, y_2, z_2)$$
$$+V(x_1, y_1, z_1)+V(x_2, y_2, z_2)=E_T$$

Since the first and third terms are only a function of (x_1, y_1, z_1) and the second and fourth terms are only a function of (x_2, y_2, z_2), and furthermore since the equation must be satisfied for independent choices of (x_1, y_1, z_1) and (x_2, y_2, z_2), it follows that we must independently

satisfy two equations namely

$$-\frac{\hbar^2}{2m} \frac{1}{\psi(x_1, y_1, z_1)} \left(\frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial y_1^2} + \frac{\partial^2}{\partial z_1^2} \right) \psi(x_1, y_1, z_1) + V(x_1, y_1, z_1) = E_1$$

and

$$-\frac{\hbar^2}{2m} \frac{1}{\psi(x_2, y_2, z_2)} \left(\frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial y_2^2} + \frac{\partial^2}{\partial z_2^2} \right) \psi(x_2, y_2, z_2) + V(x_2, y_2, z_2) = E_2$$

These are both identical one-electron Schrödinger equations. We have used the technique of separation of variables.

We have considered only the spatial parts of the wavefunctions of the electrons; however, electrons also have spin. In order to include spin the wavefunctions must also define the spin direction of the electron.

We will write a complete wavefunction $[\psi(x_1, y_1, z_1)\psi(s)]_a$, which is the wavefunction for one electron where $\psi(x_1, y_1, z_1)$ describes the spatial part and the spin wavefunction $\psi(s)$ describes the spin part, which can be spin up or spin down. There will be four quantum numbers associated with each wavefunction of which the first three arise from the spatial part. A fourth quantum number, which can be $+1/2$ or $-1/2$ for the spin part, defines the direction of the spin part. Rather than writing the full set of quantum numbers for each wavefunction we will use the subscript a to denote the set of four quantum numbers. For the other electron the analogous wavefunction is $[\psi(x_2, y_2, z_2)\psi(s)]_b$ indicating that this electron has its own set of four quantum numbers denoted by subscript b.

Now the wavefunction of the two-electron system including spin becomes

$$\psi_{T_1} = [\psi(x_1, y_1, z_1)\psi(s)]_a [\psi(x_2, y_2, z_2)\psi(s)]_b \tag{3.8a}$$

The probability distribution function, which describes the spatial probability density function of the two-electron system, is $|\psi_T|^2$, which can be written as

$$\begin{aligned}
|\psi_{T_1}|^2 &= \psi_{T_1}^* \psi_{T_1} \\
&= [\psi(x_1, y_1, z_1)\psi(s)]_a^* [\psi(x_2, y_2, z_2)\psi(s)]_b^* [\psi(x_1, y_1, z_1)\psi(s)]_a \\
&\quad \psi(x_2, y_2, z_2)\psi(s)]_b
\end{aligned} \tag{3.8b}$$

If the electrons were distinguishable then we would need also to consider the case where the electrons were in the opposite states, and in this case

$$\psi_{T_2} = [\psi(x_1, y_1, z_1)\psi(s)]_b [\psi(x_2, y_2, z_2)\psi(s)]_a \tag{3.9a}$$

Now the probability density of the two-electron system would be

$$\begin{aligned}
|\psi_{T_2}|^2 &= \psi_{T_2}^* \psi_{T_2} \\
&= [\psi(x_1, y_1, z_1)\psi(s)]_b^* [\psi(x_2, y_2, z_2)\psi(s)]_a^* [\psi(x_1, y_1, z_1)\psi(s)]_b \\
&\quad [\psi(x_2, y_2, z_2)\psi(s)]_a
\end{aligned} \tag{3.9b}$$

Clearly Equation 3.9b is not the same as Equation 3.8b, and when the subscripts are switched the form of $|\psi_T|^2$ changes. This specifically contradicts the requirement that measurable quantities such as the spatial distribution function of the two-electron system remain the same regardless of the interchange of the electrons.

In order to resolve this difficulty it is possible to write wavefunctions of the two-electron system that are linear combinations of the two possible electron wavefunctions.

We write a *symmetric* wavefunction ψ_S for the two-electron system as

$$\psi_S = \frac{1}{\sqrt{2}}[\psi_{T_1} + \psi_{T_2}] \tag{3.10}$$

and an *antisymmetric* wavefunction ψ_A for the two-electron system as

$$\psi_A = \frac{1}{\sqrt{2}}[\psi_{T_1} - \psi_{T_2}] \tag{3.11}$$

If ψ_S is used in place of ψ_T to calculate the probability density function $|\psi_S|^2$, the result will be independent of the choice of the subscripts. In addition since both ψ_{T_1} and ψ_{T_2} are valid solutions to Schrödinger's equation (Equation 3.6) and since ψ_S is a linear combination of these solutions it follows that ψ_S is also a valid solution. The same argument applies to ψ_A. See Problem 3.20.

We will now examine just the spin parts of the wavefunctions for each electron. We need to consider all possible spin wavefunctions for the two electrons. The individual electron spin wavefunctions must be multiplied to obtain the spin part of the wavefunction for the two-electron system as indicated in Equations 3.8 or 3.9, and we obtain four possibilities, namely $\psi_{\frac{1}{2}}\psi_{-\frac{1}{2}}$ or $\psi_{-\frac{1}{2}}\psi_{\frac{1}{2}}$ or $\psi_{\frac{1}{2}}\psi_{\frac{1}{2}}$ or $\psi_{-\frac{1}{2}}\psi_{-\frac{1}{2}}$.

For the first two possibilities to satisfy the requirement that the spin part of the new two-electron wavefunction does not depend on which electron is which, a symmetric or an antisymmetric spin function is required. In the symmetric case we can use a linear combination of wavefunctions

$$\psi = \frac{1}{\sqrt{2}}\left(\psi_{\frac{1}{2}}\psi_{-\frac{1}{2}} + \psi_{-\frac{1}{2}}\psi_{\frac{1}{2}}\right) \tag{3.12}$$

This is a symmetric spin wavefunction since changing the labels does not affect the result. The total spin for this symmetric system turns out to be $s = 1$. There is also an antisymmetric case for which

$$\psi = \frac{1}{\sqrt{2}}\left(\psi_{\frac{1}{2}}\psi_{-\frac{1}{2}} - \psi_{-\frac{1}{2}}\psi_{\frac{1}{2}}\right) \tag{3.13}$$

Here, changing the sign of the labels changes the sign of the linear combination but does not change any measurable properties and this is therefore also consistent with the requirements for a proper description of indistinguishable particles. In this antisymmetric system the total spin turns out to be $s = 0$.

The final two possibilities are symmetric cases since switching the labels makes no difference. These cases therefore do not require the use of linear combinations to be consistent with indistinguishability and are simply

$$\psi = \psi_{\frac{1}{2}}\psi_{\frac{1}{2}} \tag{3.14}$$

and

$$\psi = \psi_{-\frac{1}{2}}\psi_{-\frac{1}{2}} \tag{3.15}$$

These symmetric cases both have spin $s = 1$.

In summary there are four cases, three of which, given by Equations 3.12, 3.14 and 3.15, are symmetric spin states and have total spin $s = 1$, and one of which, given by Equation 3.13, is antisymmetric and has total spin $s = 0$. Note that total spin is not always simply the sum of the individual spins of the two electrons, but must take into account the addition rules for quantum spin vectors. See Eisberg et al, suggestions for further reading. The three symmetric cases are appropriately called *triplet* states and the one antisymmetric case is called a *singlet* state. Table 3.2 lists the four possible states.

We must now return to wavefunctions 3.10 and 3.11. The antisymmetric wavefunction ψ_A may be written using Equations 3.11, 3.8a and 3.9a as

$$
\begin{aligned}
\psi_A &= \frac{1}{\sqrt{2}} [\psi_{T_1} - \psi_{T_2}] \\
&= \frac{1}{\sqrt{2}} \{ [\psi(x_1, y_1, z_1)\psi(s)]_a \, [\psi(x_2, y_2, z_2)\psi(s)]_b \\
&\quad - [\psi(x_1, y_1, z_1)\psi(s)]_b \, [\psi(x_2, y_2, z_2)\psi(s)]_a \}
\end{aligned}
\tag{3.16}
$$

If, in violation of the Pauli exclusion principle, the two electrons were in the *same* quantum state $\psi_T = \psi_{T_1} = \psi_{T_2}$, which includes both position and spin, then Equation 3.16 immediately yields $\psi_A = 0$, which means that such a situation cannot occur. If the symmetric wavefunction ψ_S of Equation 3.10 was used instead of ψ_A, the value of ψ_S would not be zero for two electrons in the same quantum state. See problem 3.22. For this reason, a more complete statement of the Pauli exclusion principle is that *the wavefunction of a system of two or more indistinguishable electrons must be antisymmetric*.

In order to obtain an antisymmetric wavefunction, from Equation 3.16 either the spin part *or* the spatial part of the wavefunction may be antisymmetric. If the spin part is antisymmetric, which is a singlet state, then the Pauli exclusion restriction on the spatial part of the wavefunction may be lifted. The two electrons may occupy the same spatial wavefunction and they may have a high probability of being close to each other.

If the spin part is symmetric this is a triplet state and the spatial part of the wavefunction must be antisymmetric. The spatial density function of the antisymmetric wavefunction causes the two electrons to have a higher probability of existing further apart because they are in distinct spatial wavefunctions.

If we now introduce the coloumb repulsion between the electrons it becomes evident that if the spin state is a singlet state the repulsion will be higher because the electrons spend more time close to each other. If the spin state is a triplet state the repulsion is weaker because the electrons spend more time further apart.

Now let us return to the helium atom as an example of this. Assume one helium electron is in the ground state of helium, which is the 1s state, and the second helium electron is in an excited state. This corresponds to an excited helium atom, and we need to understand this configuration because radiation always involves excited states.

The two helium electrons can be in a triplet state or in a singlet state. Strong dipole radiation is observed from the singlet state only, and the triplet states do not radiate. We can understand the lack of radiation from the triplet states by examining spin. The total spin of a triplet state is $s = 1$. The ground state of helium, however, has no net spin because if the two electrons are in the same $n = 1$ energy level the spins must be in opposing directions to satisfy the Pauli exclusion principle, and there is no net spin. The ground state of helium

Table 3.2 *Possible spin states for a two electron system*

State	Probability	Total spin	Spin arrangement	Spin symmetry	Spatial symmetry	Spatial attributes	Dipole-allowed transition to/from singlet ground state
Singlet	25%	0	$\psi_{\frac{1}{2}}\psi_{-\frac{1}{2}} - \psi_{-\frac{1}{2}}\psi_{\frac{1}{2}}$	Antisymmetric	Symmetric	Electrons close to each other	Yes
Triplet	75%	1	$\psi_{\frac{1}{2}}\psi_{-\frac{1}{2}} + \psi_{-\frac{1}{2}}\psi_{\frac{1}{2}}$ or $\psi_{\frac{1}{2}}\psi_{\frac{1}{2}}$ or $\psi_{-\frac{1}{2}}\psi_{-\frac{1}{2}}$	Symmetric	Antisymmetric	Electrons far apart	No

is therefore a singlet state. There can be no triplet states in the ground state of the helium atom. See Problem 3.21.

There is a net magnetic moment generated by an electron due to its spin. This fundamental quantity of magnetism due to the spin of an electron is known as the *Bohr magneton*. If the two helium electrons are in a triplet state there is a net magnetic moment, which can be expressed in terms of the Bohr magneton since the total spin $s = 1$. This means that a magnetic moment exists in the excited triplet state of helium. Photons have no charge and hence no magnetic moment. Because of this a *dipole transition from an excited triplet state to the ground singlet state is forbidden* because the triplet state has a magnetic moment but the singlet state does not, and the net magnetic moment cannot be conserved. In contrast to this the *dipole transition from an excited singlet state to the ground singlet state is allowed* and strong dipole radiation is observed.

The triplet states of helium are slightly lower in energy than the singlet states. The triplet states involve symmetric spin states, which means that the spin parts of the wavefunctions are symmetric. This forces the spatial parts of the wavefunctions to be antisymmetric, as illustrated in Figure 3.8, and the electrons are, on average, more separated. As a result the repulsion between the ground state electron and the excited state electron is weaker. The excited state electron is therefore more strongly bound to the nucleus and it exists in a lower energy state. The observed radiation is consistent with the energy difference between the higher energy singlet state and the ground singlet state. As expected there is no observed radiation from the triplet excited state and the ground singlet state.

We have used helium atoms to illustrate the behaviour of a two-electron system; however, we now need to apply our understanding of these results to molecular electrons, which are important for organic light-emitting and absorbing materials. Molecules are the basis for organic electronic materials and molecules always contain two or more electrons in a molecular system.

3.7   Molecular Excitons

In inorganic semiconductors electrons and holes exist as distributed wavefunctions, which prevents the formation of stable excitons at room temperature. In contrast to this, holes and electrons are localized within a given molecule in *organic semiconductors*, and the *molecular exciton* is thereby both stabilized and bound within a molecule of the organic semiconductor. In organic semiconductors, which are composed of molecules, excitons are clearly evident at room temperature and also at higher operational device temperatures.

An exciton in an organic semiconductor is an excited state of the molecule. A molecule contains a series of electron energy levels associated with a series of *molecular orbitals* that are complicated to calculate directly from Schrödinger's equation due to the complex shapes of molecules. These molecular orbitals may be occupied or unoccupied. When a molecule absorbs a quantum of energy that corresponds to a transition from one molecular orbital to another higher energy molecular orbital, the resulting electronic excited state of the molecule is a molecular exciton comprising an electron and a hole within the molecule. An electron is said to be found in the lowest unoccupied molecular orbital and a hole in the highest occupied molecular orbital, and since they are both contained within the

(a)

(b)
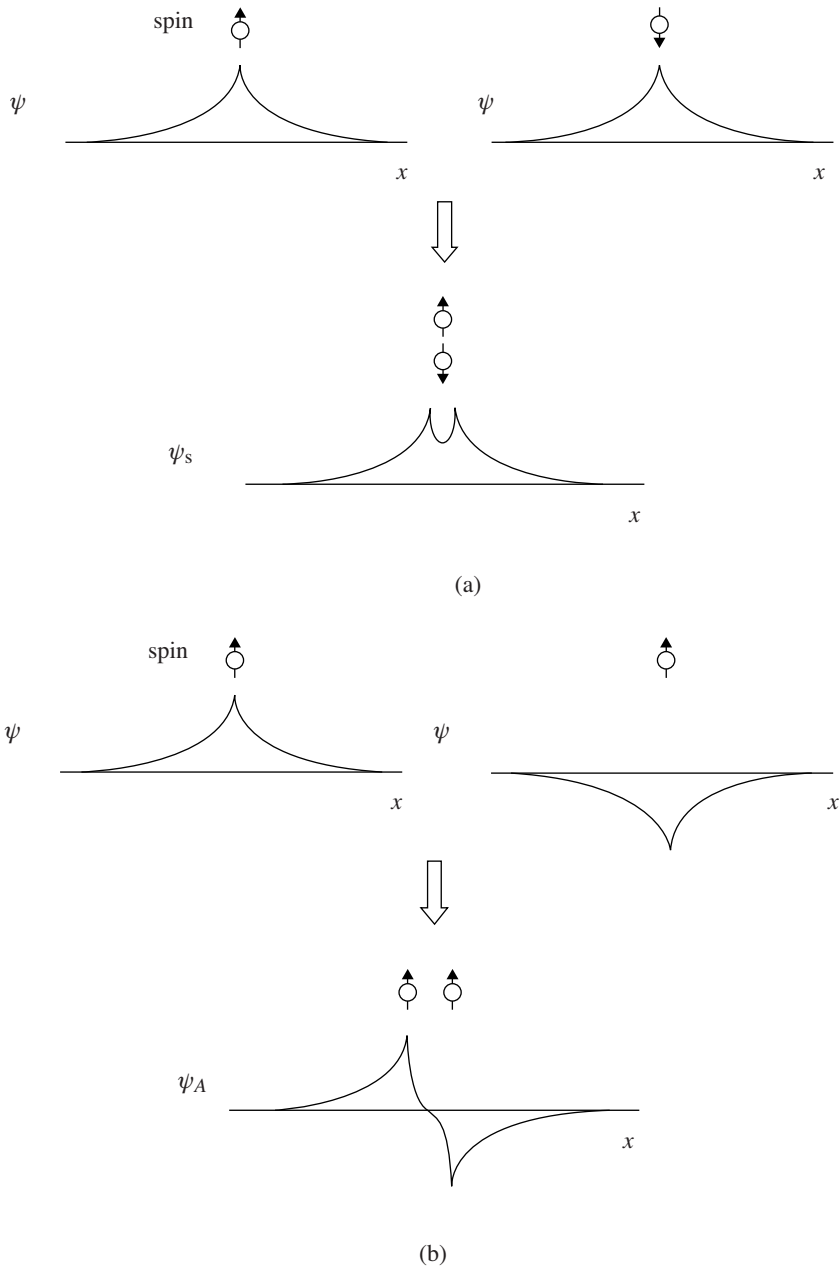
Figure 3.8   *A depiction of the symmetric and antisymmetric wavefunctions and spatial density functions of a two-electron system. (a) Singlet state with electrons closer to each other on average. (b) Triplet state with electrons further apart on average*

same molecule the electron-hole state is said to be *bound*. A *bound exciton* results, which is spatially localized to a given molecule in an organic semiconductor. Organic molecule energy levels are discussed in more detail in Chapter 6.

These molecular excitons can be classified as in the case of excited states of the helium atom, and either singlet or triplet excited states in molecules are possible. The results from Section 3.6 are relevant to these molecular excitons and the same concepts involving electron spin, the Pauli exclusion principle and indistinguishability are relevant because the molecule contains two or more electrons.

If a molecule in its unexcited state absorbs a photon of light it may be excited forming an exciton in a singlet state with spin $s = 0$. These excited molecules typically have characteristic lifetimes on the order of nanoseconds, after which the excitation energy may be released in the form of a photon and the molecule undergoes *fluorescence* by a dipole-allowed process returning to its ground state.

It is also possible for the molecule to be excited to form an exciton by electrical means rather than by the absorption of a photon. This will be described in detail in Chapter 6. Under electrical excitation the exciton may be in a singlet or a triplet state since electrical excitation, unlike photon absorption, does not require the total spin change to be zero. There is a 75% probability of a triplet exciton and 25% probability of a singlet exciton, as described in Table 3.2. The probability of fluorescence is therefore reduced under electrical excitation to 25% because the decay of triplet excitons is not dipole-allowed.

Another process may take place, however. Triplet excitons have a spin state with $s = 1$ and these spin states can frequently be coupled with the orbital angular momentum of molecular electrons, which influences the effective magnetic moment of a molecular exciton. The restriction on dipole radiation can be partly removed by this coupling, and light emission over relatively long characteristic radiation lifetimes is observed in specific molecules. These longer lifetimes from triplet states are generally on the order of milliseconds and the process is called *phosphorescence*, in contrast with the shorter lifetime fluorescence from singlet states. Since excited triplet states have slightly lower energy levels than excited singlet states, triplet phosphorescence has a longer wavelength than singlet fluorescence in a given molecule.

In addition there are other ways that a molecular exciton can lose energy. There are three possible energy loss processes that involve energy transfer from one molecule to another molecule. One important process is known as *Förster resonance energy transfer*. Here a molecular exciton in one molecule is established but a neighbouring molecule is not initially excited. The excited molecule will establish an oscillating dipole moment as its exciton starts to decay in energy as a superposition state. The radiation field from this dipole is experienced by the neighbouring molecule as an oscillating field and a superposition state in the neighbouring molecule is also established. The originally excited molecule loses energy through this resonance energy transfer process to the neighbouring molecule and finally energy is conserved since the initial excitation energy is transferred to the neighbouring molecule *without the formation of a photon*. This is not the same process as photon generation and absorption since a complete photon is never created; however, only dipole-allowed transitions from excited singlet states can participate in Förster resonance energy transfer.

Förster energy transfer depends strongly on the intermolecular spacing, and the rate of energy transfer falls off as $\frac{1}{R^6}$ where R is the distance between the two molecules. A

simplified picture of this can be obtained using the result for the electric field of a static dipole. This field falls off as $\frac{1}{R^3}$. Since the energy density in a field is proportional to the square of the field strength it follows that the energy available to the neighbouring molecule falls of as $\frac{1}{R^6}$. This then determines the rate of energy transfer.

Dexter electron transfer is a second energy transfer mechanism in which an excited electron state transfers from one molecule (the donor molecule) to a second molecule (the acceptor molecule). This requires a wavefunction overlap between the donor and acceptor, which can only occur at extremely short distances typically of the order 10–20 Å.

The Dexter process involves the transfer of the electron and hole from molecule to molecule. The donor's excited state may be exchanged in a single step, or in two separate charge exchange steps. The driving force is the decrease in system energy due to the transfer. This implies that the donor molecule and acceptor molecule are different molecules. This will be discussed in Chapter 6 in the context of organic LEDs. The Dexter energy transfer rate is proportional to $e^{-\alpha R}$ where R is the intermolecular spacing. The exponential form is due to the exponential decrease in the wavefunction density function with distance.

Finally a third process is radiative energy transfer. In this case a photon emitted by the host is absorbed by the guest molecule. The photon may be formed by dipole radiation from the host molecule and absorbed by the converse process of dipole absorption in the guest molecule. See Section 3.4.

3.8   Band-to-Band Transitions

In inorganic semiconductors the recombination between an electron and a hole occurs to yield a photon, or conversely the absorption of a photon yields a hole-electron pair. The electron is in the conduction band and the hole is in the valence band. It is very useful to analyse these processes in the context of band theory from Chapter 1.

Consider the direct-gap semiconductor having approximately parabolic conduction and valence bands near the bottom and top of these bands respectively, as in Figure 3.9. Parabolic bands were introduced in Section 1.5. Two possible transition energies, E_1 and E_2, are shown, which produce two photons having two different wavelengths. Due to the very small momentum of a photon, the recombination of an electron and a hole occurs almost vertically in this diagram to satisfy conservation of momentum. The x-axis represents the wave-number k, which is proportional to momentum. See Section 1.12.

Conduction band electrons have energy

$$E_e = E_c + \frac{\hbar^2 k^2}{2m_e^*}$$

and for holes we have

$$E_h = E_v - \frac{\hbar^2 k^2}{2m_h^*}$$

In order to determine the emission/absorption spectrum of a direct-gap semiconductor we need to find the probability of a recombination taking place as a function of energy E. This
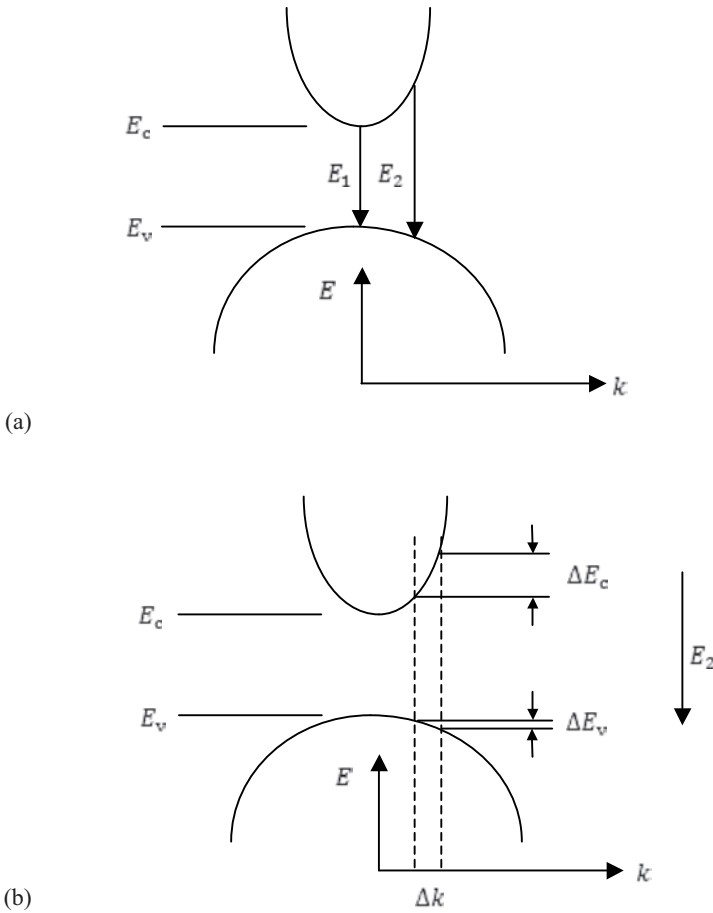
Figure 3.9   *(a) Parabolic conduction and valence bands in a direct-gap semiconductor show-ing two possible transitions. (b) Two ranges of energies ΔE_v in the valence band and ΔE_c in the conduction band determine the photon emission rate in a small energy range about a specific transition energy. Note that the two broken vertical lines in (b) show that the range of transition energies at E_2 is the sum of ΔE_c and ΔE_v*

transition probability depends on an appropriate density of states function multiplied by probability functions that describe whether or not the states are occupied.

We will first determine the appropriate density of states function. Any transition in Figure 3.9 takes place at a fixed value of reciprocal space where k is constant. The same set of points located in reciprocal space or k-space gives rise to states both in the valence band and in the conduction band. In the Kronig–Penny model presented in Chapter 1, a given position on the k-axis intersects all the energy bands including the valence and conduction bands. There is therefore a state in the conduction band corresponding to a state in the valence band at a specific value of k. Reciprocal space and k-space are equivalent except

for a constant factor of π, as explained in Section 1.9. See Figures 1.11 and 1.12, which depict reciprocal space.

Therefore in order to determine the *photon emission rate* over a specific range of photon energies we need to find the appropriate density of states function for a *transition* between a group of states in the conduction band and the corresponding group of states in the valence band. This means we need to determine the number of states in reciprocal space or k-space that give rise to the corresponding set of transition energies that can occur over a small radiation energy range ΔE centred at some transition energy in Figure 3.9. For example, the appropriate number of states can be found at E_2 in Figure 3.9b by considering a small range of k-states Δk that correspond to small differential energy ranges ΔE_c and ΔE_v and then finding the total number of band states that fall within the range ΔE. The emission energy from these states will be centred at E_2 and will have an emission energy range $\Delta E = \Delta E_\mathrm{c} + \Delta E_\mathrm{v}$ producing a portion of the observed emission spectrum. The *density of transitions* is determined by the density of states in the *joint dispersion relation*, which will now be introduced:

The available energy for any transition is given by

$$E(k) = h\upsilon = E_\mathrm{e}(k) - E_\mathrm{h}(k)$$

and upon substitution we can obtain the joint dispersion relation, which adds the dispersion relations from both the valence and conduction bands. We can express this transition energy E and determine the joint dispersion relation from Figure 3.9a as

$$E(k) = h\upsilon = E_\mathrm{c} - E_\mathrm{v} + \frac{\hbar^2 k^2}{2m_\mathrm{e}^*} + \frac{\hbar^2 k^2}{2m_\mathrm{h}^*} = E_\mathrm{g} + \frac{\hbar^2 k^2}{2\mu} \qquad (3.17)$$

where

$$\frac{1}{\mu} = \frac{1}{m_\mathrm{e}^*} + \frac{1}{m_\mathrm{h}^*}$$

Note that a range of k-states Δk will result in an energy range $\Delta E = \Delta E_\mathrm{c} + \Delta E_\mathrm{v}$ in the joint dispersion relation because the joint dispersion relation provides the *sum* of the relevant ranges of energy in the two bands as required. The smallest possible value of transition energy E in the joint dispersion relation occurs at $k = 0$ where $E = E_\mathrm{g}$ from Equation 3.17, which is consistent with Figure 3.9. If we can determine the density of states in the joint dispersion relation we will therefore have the density of possible photon emission transitions available in a certain range of energies.

The density of states function for an energy band was derived in Section 1.9. As originally derived, the form of the density of states function was valid for a box having $V = 0$ inside the box. In an energy band, however, the density of states function was modified. We replaced the free electron mass with the reduced mass relevant to the specific energy band and we replaced m in Equation 1.23 by m^* to obtain Equation 1.31a. This is valid because rather than the parabolic E versus k dispersion relation for free electrons in which the electron mass is m, we used the parabolic E versus k dispersion relation for an electron in an energy band as illustrated in Figure 1.7, which may be approximated as parabolic for small values of k with the appropriate effective mass. The slope of the E versus k dispersion relation is controlled by the effective mass, and this slope determines density of states

along the energy axis for a given density of states along the k-axis. We can now use the same method to determine the density of states in the joint dispersion relation of Equation 3.17 by substituting the reduced mass μ into Equation 1.23. Recognizing that the density of states function must be zero for $E < E_g$ we obtain

$$D_{\text{joint}}(E) = \frac{1}{2}\pi \left(\frac{2\mu}{\pi^2 \hbar^2}\right)^{\frac{3}{2}} (E - E_g)^{\frac{1}{2}} \tag{3.18}$$

This is known as the *joint density of states* function valid for $E \geq E_g$.

To determine the probability of occupancy of states in the bands, we use Fermi–Dirac statistics, introduced in Chapter 1. The Boltzmann approximation for the probability of occupancy of carriers in a conduction band was previously obtained in Equation 1.33 as

$$F(E) \cong \exp\left[-\frac{(E_e - E_f)}{kT}\right]$$

and for a valence band the probability of a hole is given by

$$1 - F(E) \cong \exp\left[\frac{(E_h - E_f)}{kT}\right]$$

Since a transition requires both an electron in the conduction band and a hole in the valence band, the probability of a transition will be proportional to

$$F(E)[1 - F(E)] = \exp\left[-\frac{(E_e - E_h)}{kT}\right] = \exp\left[-\frac{E}{kT}\right] \tag{3.19}$$

Including the density of states function, we conclude that the probability $p(E)$ of an electron-hole pair recombination applicable to an LED is proportional to the product of the joint density of states function and the function $F(E)[1 - F(E)]$, which yields

$$p(E) \propto D(E - E_g)F(E)[1 - F(E)] \tag{3.20}$$

Now using Equations 3.18, 3.19 and 3.20 we obtain the photon emission rate $R(E)$ as

$$R(E) \propto (E - E_g)^{1/2} \exp\left[-\frac{E}{kT}\right] \tag{3.21}$$

The result is shown graphically in Figure 3.10.

It is important to note that the result in Equation 3.21 is important for direct gap semiconductors but differs fundamentally from the recombination rate $R \propto np$ previously presented in Section 1.15. This is because in Section 1.15 we assumed that carriers scattered and did not maintain a specific value of k for recombination, which is particularly relevant in an indirect gap semiconductor such as silicon. The recombination events in a semiconductor such as silicon are more likely to be non-radiative than radiative and phonon interactions are involved. In this case the overall carrier concentrations n and p in the bands are important and the consideration of specific values of k used to derive Equation 3.21 are not relevant.

If we differentiate Equation 3.21 with respect to E and set $\dfrac{dR(E)}{dE} = 0$ the maximum is found to occur at $E = E_g + \dfrac{1}{2kT}$. From this, we can evaluate the full width at half
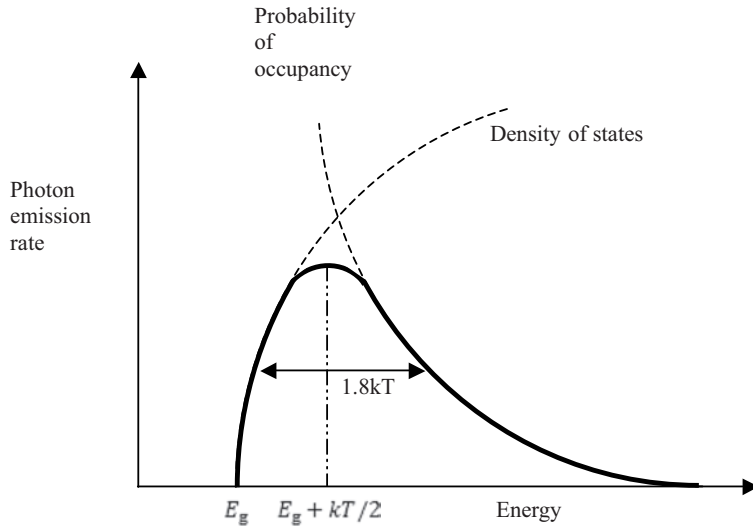
Figure 3.10 *Photon emission rate as a function of energy for a direct gap transition of an LED. Note that at low energies the emission drops off due to the decrease in the density of states term $(E - E_g)^{\frac{1}{2}}$ and at high energies the emission drops off due to the Boltzmann term $\exp\left[-\dfrac{E}{kT}\right]$*

maximum to be $1.8kT$. See Problem 3.18. This will be further discussed in Chapter 5 in the context of LEDs.

For a solar cell, the absorption constant α can be evaluated using Equation 3.6, and we obtain

$$\alpha(h\nu) \propto (h\nu - E_g)^{\frac{1}{2}} \tag{3.22}$$

We consider the valence band to be fully occupied by electrons and the conduction band to be empty. In this case the absorption rate depends on the joint density of states function only and is independent of Fermi–Dirac statistics. The absorption edge for a direct-gap semiconductor is illustrated in Figure 3.11.

This absorption edge is only valid for direct-gap semiconductors, and only when parabolic band-shapes are valid. If $h\nu \gg E_g$ this will not be the case and measured absorption coefficients will differ from this theory.

In an indirect gap semiconductor the absorption α increases more gradually with photon energy $h\nu$ until a direct-gap transition can occur. This is discussed in more detail in Chapter 4 in the context of solar cells.

3.9   Photometric Units

The most important applications of light emitting diodes (LEDs) – to be described in Chapter 5 – and organic light emitting diodes (OLEDs) – to be described in Chapter 6 – are for visible illumination and displays. This requires the use of units to measure the
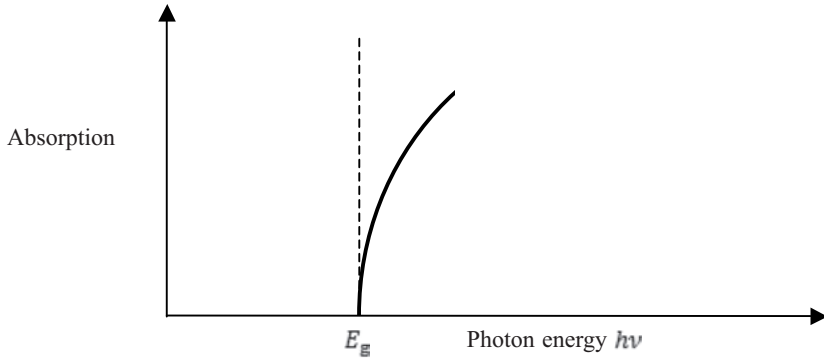
Figure 3.11   *Absorption edge for direct-gap semiconductor*

brightness and colour of light output. The power in watts and wavelength of emission are often not adequate descriptors of light emission. The human visual system has a variety of attributes that have given rise to more appropriate units and ways of measuring light output. This human visual system includes the eye, the optic nerve and the brain, which interpret light in a unique way. Watts, for example, are considered *radiometric* units, and this section introduces *photometric* units and relates them to radiometric units.

Luminous intensity is a photometric quantity that represents the perceived brightness of an optical source by the human eye. The unit of luminous intensity is the *candela* (cd). One cd is the luminous intensity of a source that emits 1/683 watt of light at 555 nm into a solid angle of one steradian. The candle was the inspiration for this unit, and a candle does produce a luminous intensity of approximately one candela.

Luminous flux is another photometric unit that represents the light power of a source. The unit of luminous flux is the *lumen* (lm). A candle that produces a luminous intensity of 1 cd produces 4π lumens of light power. If the source is spherically symmetrical then there are 4π steradians in a sphere, and a luminous flux of 1 lm is emitted per steradian.

A third quantity, *luminance*, refers to the luminous intensity of a source divided by an area through which the source light is being emitted; it has units of cd m^{-2}. In the case of an LED *die* or semiconductor chip light source the luminance depends on the size of the die. The smaller the die that can achieve a specified luminous intensity, the higher the luminance of this die.

The advantage of these units is that they directly relate to perceived brightnesses, whereas radiation measured in watts may be visible, or invisible depending on the emission spectrum. Photometric units of luminous intensity, luminous flux and luminance take into account the relative sensitivity of the human vision system to the specific light spectrum associated with a given light source.

The eye sensitivity function is well known for the average human eye. Figure 3.12 shows the perceived brightness for the human visual system of a light source that emits a constant optical power that is independent of wavelength. The left scale has a maximum of 1 and is referenced to the peak of the human eye response at 555 nm. The right scale is in units of *luminous efficacy* (lm W^{-1}), which reaches a maximum of 683 lm W^{-1} at 555 nm. Using Figure 3.12, luminous intensity can now be determined for other wavelengths of light.
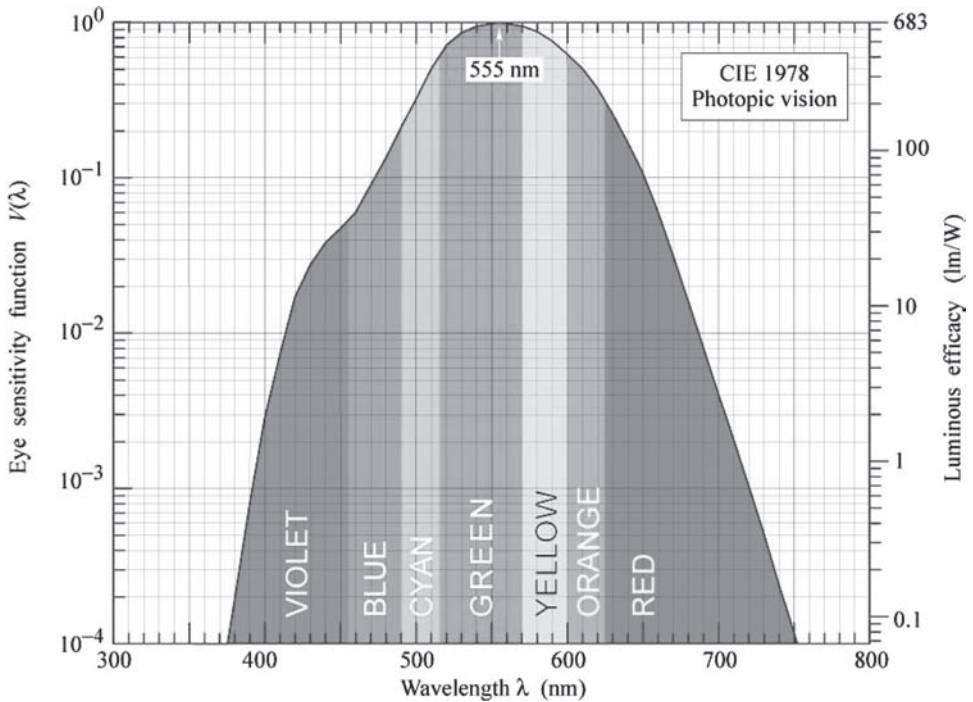
Figure 3.12 *The eye sensitivity function. The left scale is referenced to the peak of the human eye response at 555 nm. The right scale is in units of luminous efficacy. International Commission on Illumination (Commission Internationale de l'Eclairage, or CIE), 1931*

An important measure of the overall efficiency of a light source can be obtained using luminous efficacy from Figure 3.12. A hypothetical monochromatic electroluminescent light source emitting at 555 nm that consumes 1 W of electrical power and produces 683 lm has an electrical-to-optical conversion efficiency of 100%. A hypothetical monochromatic light source emitting at 410 nm that consumes 1 W of electrical power and produces approximately 5 lm also has a conversion efficiency of 100%. The luminous flux of a blue LED or a red LED that consumes 1 W of electric power may be lower than for a green LED; however, this does not necessarily mean that they are less efficient.

Luminous efficiency values for a number of light sources may be described in units of lm W^{-1}, or light power divided by electrical input power. Table 3.3 adds the luminous efficiency values relevant to Table 3.1. Luminous efficiency can never exceed luminous efficacy for a light source having a given spectrum.

The perceived colour of a light source is determined by its spectrum. The human visual system and the brain create our perception of colour. For example, we often perceive a mixture of red and green light as yellow even though none of the photons arriving at our eyes is yellow. For a LED, both the peak and the full width at half-maximum (FWHM) of the emission spectrum determine the colour. For a down-converted spectrum, as shown in Figure 5.23, the perceived colour is a function of the entire spectrum including both the blue peak and the yellow peak.

Table 3.3 *Luminous efficiency values (in lm W^{-1}) for a variety of light emitters*

Blackbody radiation (light generated due to the temperature of a body)	Sun Tungsten filament lamp: $\eta \cong 5\%$ or 15–20 lm W^{-1}
Photoluminescence (light emitted by a material that is stimulated by electromagnetic radiation)	Fluorescent lamp phosphors ($\eta \cong 80\%$) The fluorescent lamp achieves 50–80 lm W^{-1} and this includes the generation of UV in a gas discharge
Cathodoluminescence (light emitted by a material that is excited by high-energy electrons incident from a vacuum)	Cathode-ray tube screen ($\eta \cong 20\%$) A green-emitting cathode-ray tube achieves approximately 50 lm W^{-1}. Colour tubes have considerably lower efficiencies of 1–2 lm W^{-1} due to the use of a shadow mask that drains away about 80% of the beam current
Electroluminescence (light emitted by a material that is directly electrically excited)	Light emitting diode ($\eta \cong 20 - 50\%$). For a 20% efficient LED this translates into the following approximate values: red LED at 625 nm 40 lm W^{-1} green LED at 530 nm 120 lm W^{-1} blue LED at 470 nm 12 lm W^{-1}

The human eye contains light receptors on the retina that are sensitive in fairly broad bands centred at the red, the green and the blue parts of the visible spectrum. Colour is determined by the relative stimulation of these receptors. For example, a light source consisting of a combination of red and green light excites the red and green receptors, as does a pure yellow light source, and we therefore perceive both light sources as yellow in colour.

Since the colours we observe are perceptions of the human visual system, a colour space has been developed and formalized that allows all the colours we recognize to be represented on a two-dimensional graph called the *colour space chromaticity diagram* (Figure 3.13). The diagram was created by the International Commission on Illumination (Commission Internationale de l'Éclairage, or CIE) in 1931, and is therefore often referred to as the CIE diagram. CIE *x* and *y colour coordinates* are shown that can be used to specify the colour point of any light source. The outer boundary of this colour space refers to monochromatic light sources that emit light at a single wavelength. As we move to the centre of the diagram to approach white light the light source becomes increasingly less monochromatic. Hence a source having a spectrum of a finite width will be situated some distance inside the boundary of the colour space.

If two light sources emit light at two distinct wavelengths anywhere on the CIE diagram and these light sources are combined into a single light beam, the human eye will interpret the colour of the light beam as existing on a straight line connecting the locations of the two sources on the CIE diagram. The position on the straight line of this new colour will depend on the relative radiation power from each of the two light sources. See problem 3.23.

If three light sources emit light at three distinct wavelengths that are anywhere on the CIE diagram and these light sources are combined into a single light beam, the human eye
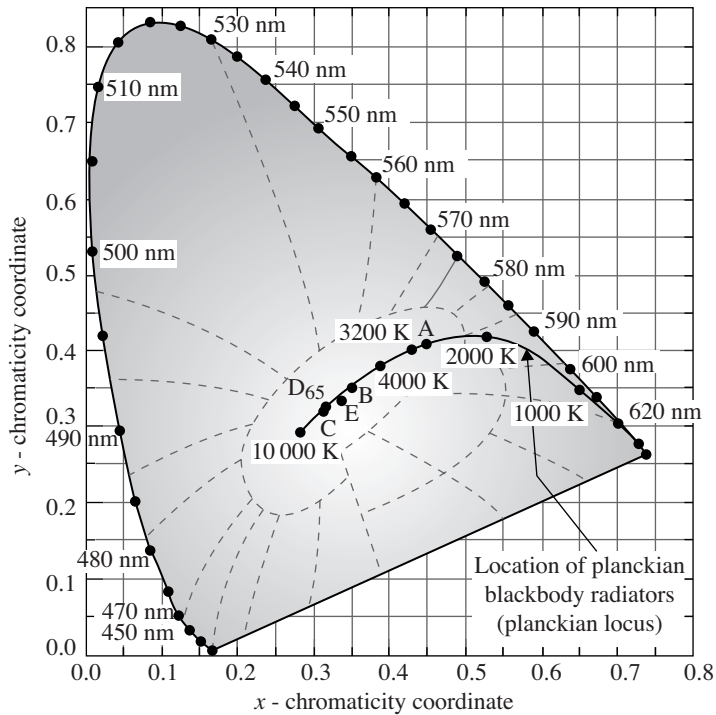
Figure 3.13   *Colour space chromaticity diagram showing colours perceptible to the human eye. The centre region of the diagram indicates a planckian locus, which corresponds to the colours of emission from a blackbody source having temperatures from 1000 K to 10 000 K. This locus includes the solar spectrum corresponding to a 5250 K blackbody. International Commission on Illumination (Commission Internationale de l'Eclairage, or CIE), 1931*

will interpret the colour of the light beam as existing within a triangular region of the CIE diagram having vertices at each of the three sources. The position within the triangle of this new colour will depend on the relative radiation power from each of the three light sources. This ability to produce a large number of colours of light from only three light sources forms the basis for *trichromatic* illumination. Lamps and displays routinely take advantage of this principle. It is clear that the biggest triangle will be available if red, green and blue light sources are selected to define the vertices of the colour triangle. This colour triangle is often referred to as a *colour space* that is enabled by a specific set of three light emitters.

3.10   Summary

3.1. Luminescence is created by accelerating charges, and examples of luminescence include blackbody radiation, photoluminescence, cathodoluminescence and electroluminescence.

3.2. Accelerating charges emit energy through the Poynting vector, which carries electric and magnetic field energy. The total power radiated from an accelerating charge is found by integrating the radiated power over a sphere and is found to be proportional to the charge and to the square of the acceleration.

3.3. The simplest mode of charge acceleration is the oscillating dipole radiator in which a charge oscillates sinusoidally. The average power radiated by the dipole is calculated by performing a time average of the instantaneous power.

3.4. Stationary quantum states do not radiate whereas superposition states may radiate through dipole radiation. The expected value of the amplitude of the oscillation of the charge is determined by $\langle \phi_n | r | \phi_{n'} \rangle$, which may be calculated. If $\langle \phi_n | r | \phi_{n'} \rangle = 0$, then radiation will not occur and the transition is forbidden.

3.5. The exciton is formed by a hole and an electron that form a hydrogen-like entity. Excitons in semiconductors give rise to absorption or emission lines that are observable at low temperatures. These lines exist inside the energy gap of the semiconductor. Excitons may either be free to travel through the semiconductor or they might be bound to a defect.

3.6. The two-electron atom involves the consideration of indistinguishable electrons. The wave functions for the two-electron atom describe either symmetric or antisymmetric states. The resulting states are known as singlet states in which the electrons are relatively close to each other; dipole radiation is associated with singlet states and not triplet states.

3.7. The molecular exciton comprises an electron and a hole that exist within one molecule. The exciton is bound within the molecule. The molecular exciton can be understood based on the two-electron atom and a singlet or a triplet exciton can be achieved. Singlet molecular excitons involve relatively fast dipole-allowed photon absorption or emission (fluorescence); however, triplet molecular excitons that may be formed by the electrical excitation of molecules decay relatively slowly (phosphorescence).

3.8. Band-to-band transitions in a direct-gap semiconductor produce a range of wavelengths depending on the position in the band. The distribution of electrons and holes in a band as a function of momentum may be determined by the density of states in the bands and the probabilities of occupancy of these states in the bands allowing the radiation spectrum of such a transition to be determined. In addition the absorption spectrum in a direct gap semiconductor can be determined.

3.9. The human eye perceives visible light in conjunction with the human brain, and a set of photometric units has been developed that allows our perception of brightness and colour to be quantified. Units of luminance in candelas per metre squared and colour coordinates (x,y) in a two-dimensional space form the basis for these units.

Suggestions for Further Reading

Ashcroft NW and Mermin ND. *Solid State Physics*. Holt, Rinehart and Winston, 1976.

Eisberg R and Resnick R. *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd edn. Wiley, 1985.

Kittel C. *Introduction to Solid State Physics*, 8th edn. Wiley, 2005.

Schubert EF. *Light Emitting Diodes*, 2nd edn. Cambridge University Press, 2006.

Problems

3.1 Consider an electron that oscillates in a dipole radiator and produces a radiation power of 1×10^{-12} watts. Find the amplitude of the oscillation of the electron if the radiation is at the following wavelengths:
(a) 470 nm (blue)
(b) 530 nm (green)
(c) 630 nm (red).

3.2 Find the time required to produce one photon from the radiator of Problem 3.1 if the radiation is:
(a) 470 nm
(b) 530 nm
(c) 630 nm.

3.3 Find the number of photons per second required to produce an optical power of 1 watt if the radiation is:
(a) 470 nm
(b) 530 nm
(c) 630 nm.

3.4 Find the number of photons per second required to produce a luminous flux of one lumen if the radiation is:
(a) 470 nm
(b) 530 nm
(c) 630 nm.

3.5 Determine the luminous efficacy and the colour coordinates of light sources that emit at a single wavelength at:
(a) 470 nm
(b) 530 nm
(c) 630 nm.

3.6 Find the electric field $\varepsilon_{\perp}$ generated a distance of 100 nm from an electron that accelerates at 10 000 m s^{-2}. Plot your result as a function of the angle between the acceleration vector and the line joining the electron to the point at which the electric field is measured. Repeat for the magnetic field $B_{\perp}$. Find the magnitude using appropriate units and the direction of the Poynting vector for the resulting travelling wave.

3.7 Find the total average power radiated from the charge of Problem 3.6.

3.8 Find the total average power radiated from an electron that oscillates at:
(a) A frequency of 10^{14} Hz and amplitude of oscillation of 0.2 nm.
(b) A frequency of 10^{15} Hz and amplitude of oscillation of 0.2 nm.
(c) A frequency of 10^{15} Hz and amplitude of oscillation of 0.5 nm.
(d) In what part of the electromagnetic spectrum will the radiation be for (a) and (b)? Find the wavelength in free space for the radiation of (a) and (b).

3.9 For each of (a), (b) and (c) in Problem 3.8, find the length of time needed to produce one photon. What is the photon energy? What is the photon emission rate in photons per second? How many oscillations of the electron are required to produce one phonon for each case?

3.10 Find the energy of an exciton that is formed from an electron and a hole in gallium arsenide using the appropriate effective masses and relative dielectric constant. Repeat for gallium nitride and cadmium sulphide.

3.11 Find the radius of an exciton in its ground state for each of the semiconductors in Problem 3.10.

3.12 Using a computer, plot the photon emission rate $R(E)$ as a function of energy for a band-to-band recombination event in a direct-gap semiconductor at room temperature. Cover an energy range that extends $10kT$ below and $10kT$ above the bandgap. Assume a bandgap of 2 eV.

3.13 Silicon, being an indirect gap semiconductor, is very inefficient as an emission source from band-to-band radiation. Nevertheless if highly purified and extremely low defect density silicon is prepared then carrier lifetimes can become long enough for radiative recombination to compete effectively with non-radiative emission. Search for information on this topic and prepare a short (2–3page) report on the state of the art on this topic. Some aspects of this topic will touch on LEDs, which are covered in Chapter 5. Keywords to consider: radiative emission in silicon; the silicon light emitting diode.

3.14 The colour space defined by three light emitters to form a trichromatic system in television is an important specification that contributes to the quality of the display. Find the colour coordinates of red, green and blue emitters that are commercially accepted standards for television in both North America and in Europe and plot these on a CIE diagram. Show the correct name for each standard. Show the triangular trichromatic colour space on the CIE diagram for each standard and comment on the limitation of this space in terms of all possible colours that the human visual system can interpret. Compare your answers to the answer to Problem 3.5. Comment on the colour spaces and give possible reasons for them to be limited in size if you consider the eye sensitivity function.
Keywords to consider: CIE diagram; RGB colour coordinates; NTSC standards; PAL standards.

3.15 The colour coordinates of displays for portable electronics such as laptop computers generally provide smaller colour spaces than for television. Battery power is a critical limitation on the light sources used for the display and maximum display brightness is desired. Explain why a reduction in colour space is helpful with reference to the eye sensitivity function and the CIE diagram. See if you can obtain the colour spaces used in portable electronics.
Keywords to consider: reduced colour coordinates; portable electronics; laptop displays.

3.16 Luminance of light sources and displays varies according to the application. For the following, find the luminance levels in units of cd m^{-2} that have become standard in the industry:
(a) Electroluminescent night light
(b) Cell-phone display
(c) Movie screen in movie theatre
(d) Laptop display
(e) Desktop computer monitor
(f) Television
(g) Display for outdoor use in shop window
(h) Outdoor electronic billboard.

3.17 The light output from small area sources that approximate a directed point source, such as an inorganic light emitting diode, is specified in terms of a plot of luminous intensity as a function of angle of emission θ, zero degrees corresponding to the central axis of the emission cone. The luminous intensity of the device is generally quoted along the central axis of the emission cone.
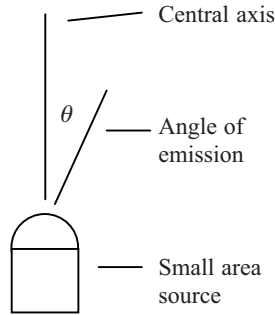
(a) Find a manufacturer's data sheet of the light spread from a high-efficiency red LED light source specified as a 30 degree device and plot the luminous intensity as a function of angle of emission using units of candelas. Your plot should cover the angle range from $\theta = -30°$ to $\theta = +30°$. The value on your graph at $\theta = 0$ should correspond to the LED's quoted luminous intensity.

(b) Use the plot from (a) and integrate the total light output from the red LED to obtain the luminous flux in units of lumens. This luminous flux represents the total light output from the LED.
Hint: The emission pattern is circular and you must use a spherical-polar coordinate system to perform this integral correctly. The area under the plot of luminous intensity as a function of angle is NOT the correct answer.

(c) Refer to the test conditions used to obtain the quoted luminous intensity for the LED of (a). Using typical values of voltage and current quoted by the manufacturer, calculate the electrical power in watts used by the device. Now divide the result of (b) by the electrical power in watts to obtain the LED efficiency. This efficiency is quoted in lumens per watt.

(d) Obtain the wavelength of emission of the red LED. Refer to Figure 3.12 and determine the luminous efficacy of the LED. Note that luminous efficacy is not the same as luminous efficiency. Divide the luminous efficiency of (c) by the luminous efficacy. This unitless quantity is the power efficiency of the LED and is a measure of the fraction of electrical input power that gets converted to light. By way of reference, high-efficiency red LEDs can achieve a power efficiency of approximately 50%. See Chapter 5.

(e) Repeat (a) to (d) for high-efficiency green and blue LEDs. Compare your results with data in Chapter 5.

3.18 Differentiate Equation 3.9 with respect to E and set $\dfrac{dR(E)}{dE} = 0$. Show that the maximum will occur at $E = E_g + \dfrac{1}{2kT}$. From this, show that the full width at half maximum (FWHM) of the LED emission spectrum is $\cong 1.8kT$.

3.19 Integrate

$$P = \frac{1}{16\pi\epsilon_0} \frac{2q^2a^2}{c^3} \int_0^\pi \sin^3\theta \, d\theta$$

to obtain Equation 3.1.

3.20 Show that both the symmetric wavefunction ψ_S and the antisymmetric wavefunction ψ_A of Equations 3.10 and 3.11, respectively, will yield probability density functions that are not in any way affected by the labelling of the two electrons.

3.21 Explain why the ground ($n = 1$) state of an atom containing two electrons can only exist as a singlet state and the triplet state cannot occur.

Hint: Consider the options available for both the spatial and the spin portions of the wavefunction to ensure that labelling of the two electrons is consistent with the requirement that the electrons are indistinguishable.

3.22 Find ψ_S written using Equations 3.10, 3.8a and 3.9a. Show that for two electrons in the same quantum state that ψ_S does not become zero as required by the Pauli Exclusion Principle.

3.23 Consider a red light source having colour coordinates $x = 0.35$, $y = 0.65$ and a green source having colour coordinates $x = 0.4$, $y = 0.6$.

(a) If the two light sources are combined on a screen such that the screen luminance due to the red source is 50 cd m^{-2} and the screen luminance due to the green source is 50 cd m^{-2}, find the resulting colour coordinates of the combined light at the screen.

Hint: Use Figure 3.12 to determine the ratio of the radiation power from each source. Plot the two colour coordinates on the colour space chromaticity diagram, and determine the colour coordinates of the resulting colour by dividing the line into two parts with lengths of this ratio.

(b) Repeat (a), but now add a third blue light source having colour coordinates $x = 0.15$, $y = 0.1$ and screen luminance 20 cd m^{-2}.

4

The Solar Cell

Objectives

1. Understand basic solar cell p-n junction design and its application to the PV industry.
2. Understand the solar spectrum as well as light absorption in semiconductors.

3. Introduce the model for the p-n junction when functioning as a solar cell and obtain an understanding of the cell efficiency as a function of design parameters.
4. Derive a description of the thin solar cell as well as an understanding of its relevance to solar cell design.
5. Describe the technology of crystalline silicon solar cell technology including silicon purification and silicon wafer production.
6. Understand the concepts of advanced silicon solar cell production including silicon ribbon technology.
7. Introduce the amorphous thin-film silicon solar cell.
8. Introduce telluride/selenide/sulphide-based thin-film solar cells as well as their structure and performance.
9. Introduce the concept of multi-junction high-efficiency solar cells.
10. Discuss concentrator solar cell advantages and limitations.

4.1 Introduction

We are now in a position to focus specifically on p-n junctions designed as solar cells for *photovoltaic* (PV) electricity production. Radiation from the sun enters the semiconductor p-n junction and generates electron-hole pairs, which are then able to flow through an external circuit and provide electrical power.

Solar cells may be small, such as the solar cells in rechargeable pocket calculators; however, the more important large-scale deployment of solar cells for electricity production is now underway and as a result the PV industry is experiencing very high growth rates. Since 2002 photovoltaic production has been increasing by an average of more than 20% annually. At the end of 2009 the cumulative global PV installation surpassed 21 GW. Photovoltaics has become the fastest growing sector of the energy production industry. The global solar cell market is expected to expand to 16.9 GW in 2012, up from 7.7 GW in 2009.

Bulk crystalline silicon solar panels currently dominate the market; however, thin-film solar cells are growing in importance and are gradually taking an increasing proportion of the global installed capacity. The thin-film portion of the installed capacity is expected to reach over 30% by 2013. Costs are dropping steadily and solar cells may now be produced at a cost of under US$1.00 per watt.

In this chapter we will study the basic operation of p-n junctions specifically designed and optimized for solar cells. Since silicon is the most important PV material we will focus on silicon; however, other materials will be introduced also. The physics of the p-n junction solar cell is relevant to all semiconductor materials and is presented first in this chapter. We will start with a qualitative band picture of the solar cell and then examine optical absorption, solar radiation and solar cell performance in more detail.

The solar cell functions as a forward-biased p-n junction; however, current flow occurs in a direction opposite to that shown in Figure 2.7. This is illustrated in Figure 4.1. Light that enters the p-n junction and reaches the depletion region of the solar cell generates
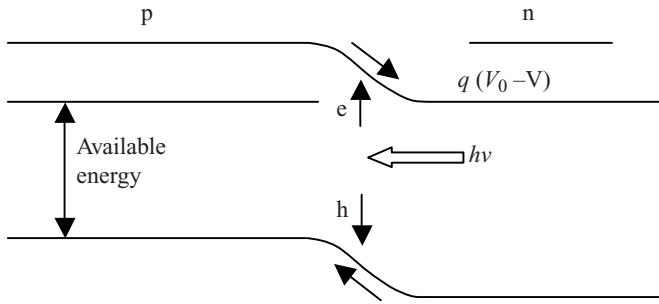
Figure 4.1 *Band diagram of a solar cell showing the directions of carrier flow. Generated electron-hole pairs drift across the depletion region. e, electron; h, hole*

electron-hole pairs. The generated minority carriers will *drift* across the depletion region and enter the n- and p-regions as majority carriers as shown. It is also possible for electron-hole pairs to be generated within about one diffusion length on either side of the depletion regions and through diffusion to reach the depletion region, where drift will again allow these carriers to cross to the opposite side. It is crucial to minimize carrier recombination, allowing carriers to deliver the available energy to the external circuit. Recombination results in heat generation. This means that the carriers must cross the depletion region and become majority carriers on the opposite side of the junction. If they are generated and recombine on the same side of the junction they will not contribute to the flow of current.

The available energy may also be optimized by minimizing the potential barrier $q(V_0 - V)$ that is required to facilitate carrier drift across the depletion region. The magnitude of $q(V_0 - V)$ is subtracted from the semiconductor bandgap, which reduces the available energy difference between electrons and holes travelling in the n-type and p-type semiconductors respectively. This causes a reduction in the operating voltage of the solar cell. If $q(V_0 - V)$ is too small, however, carriers will not be swept across the depletion region effectively enough making them more susceptible to recombination.

A useful way to think about solar cell operation is as follows: A p-n junction exhibits current–voltage behaviour as in Figure 2.7. If the p-n junction is illuminated in the junction region then in reverse bias the reverse current increases substantially due to the electron-hole pairs that are optically generated. Without optical generation, the available electrons and holes that comprise reverse saturation current are thermally generated minority carriers, which are low in concentration.

In forward bias the reverse current still flows but it is usually *smaller* than diffusion current. In a solar cell, however, the optically generated current is much *larger* than diffusion current and it continues to dominate current flow until stronger forward bias conditions are present. The *I–V* characteristic of a solar cell is shown in Figure 4.2. The appropriate operating point for a solar cell is shown in which current flows out the positive terminal of the p-n junction (p-side), through the external circuit, and then into the negative (n-side) terminal. At this operating point the current flow is still dominated by optically generated carrier drift rather than by majority carrier diffusion.

A photodiode is a light detector that operates in reverse bias, as shown in Figure 4.2. In this case, current flow is in the same direction as for solar cells, but energy is consumed
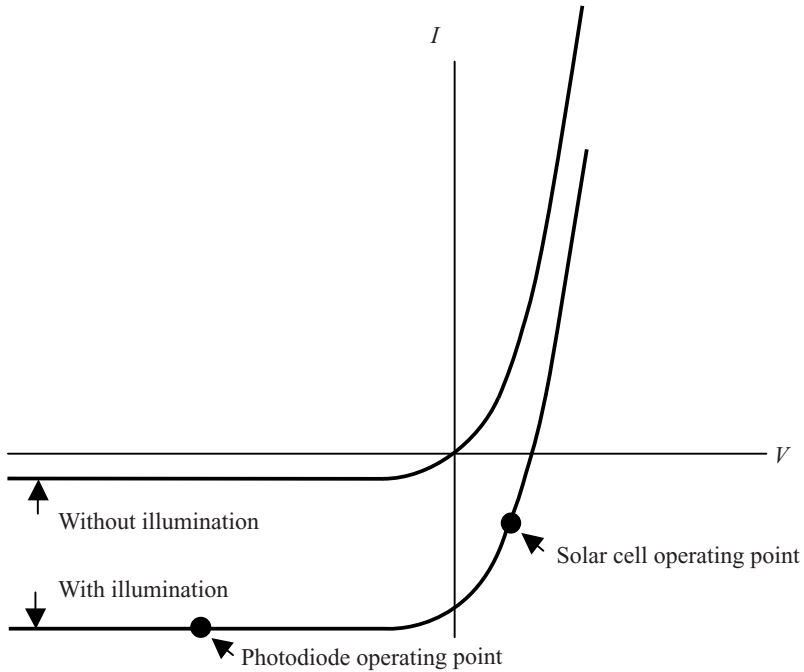
Figure 4.2   *The I–V characteristic of a solar cell or photodiode without and with illumination. The increase in reverse current occurs due to optically generated electron-hole pairs that are swept across the depletion region to become majority carriers on either side of the diode*

rather than generated because the voltage is of the opposite polarity. Photodiodes are closely related to solar cells in spite of their different mode of operation.

4.2   Light Absorption

In order to efficiently generate electron-hole pairs in a solar cell, light must reach the junction area and be absorbed effectively. Total energy will be conserved during the absorption process. Photons have energy $E = h\nu$, which must be at least as large as the semiconductor bandgap.

Total momentum will also be conserved. Photon momentum $p_{\mathrm{photon}} = \dfrac{h}{\lambda}$ is very small compared to electron and hole momentum values in typical semiconductors. From Figure 1.7, for example, the band states range in momentum $p = \hbar k$ from $p = 0$ to $p = \dfrac{\hbar \pi}{a}$. Since the lattice constant a is generally in the range of a few angstroms whereas visible light has wavelengths λ in the range of 5000 angstroms, it is clear that p_{photon} is much smaller than the possible electron and hole momentum values in a band. The absorption coefficient for a given photon energy therefore proceeds as an almost vertical transition, as illustrated in Figure 4.3. This was discussed in detail in Chapter 3.
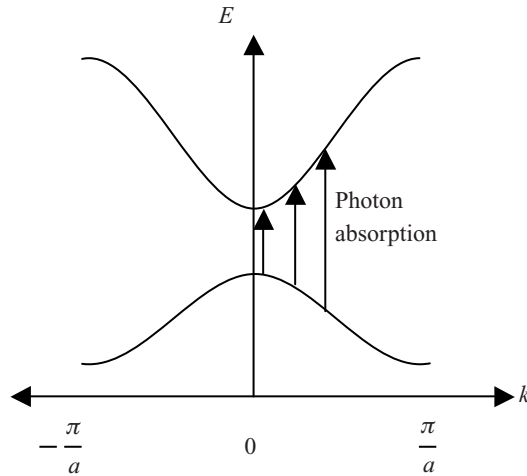
Figure 4.3 *The absorption of a photon in a direct-gap semiconductor proceeds in an almost vertical line since the photon momentum is very small on the scale of the band diagram. There are many possible vertical lines that may represent electron-hole generation by photon absorption as shown. These may exceed the bandgap energy*

Photon absorption for a direct-gap semiconductor was obtained in Equation 3.22, and therefore

$$\alpha(h\nu) \approx A(h\nu - E_\mathrm{g})^{1/2}$$

where A is a constant that depends on the material. Examples of direct-gap semiconductors used in solar cells are shown in Figure 1.13. These include the III-V semiconductor gallium arsenide and the II-VI semiconductor cadmium telluride. Several other important direct-gap solar cell semiconductor materials will be discussed later in this chapter.

In indirect gap semiconductors, the absorption of a photon having energy $h\nu \approx E_\mathrm{g}$ would appear to be forbidden due to the requirement of momentum conservation, illustrated in Figure 4.4 and discussed in Section 1.12. Absorption is possible, however, if phonons are available to supply the necessary momentum. Typical phonons in crystalline materials can transfer large values of instantaneous momentum to an electron since atomic mass is much higher than electron mass. Phonon energies are small, however. At temperature T the phonon energy will be of the order of kT, or only 0.026 eV at room temperature. The absorption process involving a phonon is a two-step process, as shown in Figure 4.4, along with a single-step absorption for higher energy photons. The result is a low but steadily increasing absorption coefficient as photon energy increases above E_g followed by a much steeper increase in absorption once photon energies are high enough for direct absorption.

The absorption coefficients for a number of important semiconductors are shown in Figure 4.5. Note the long absorption tails for silicon and germanium, which are indirect gap materials. The other semiconductors are direct-gap materials that exhibit much sharper absorption edges.
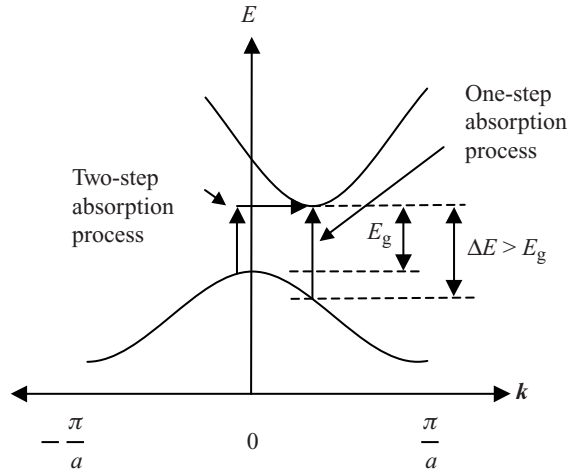
Figure 4.4   *Indirect gap semiconductor showing that absorption near the energy gap is only possible if a process involving phonon momentum is available to permit momentum conservation. The indirect, two-step absorption process involves a phonon to supply the momentum shift that is necessary to absorb photons. For higher energy photons a direct one-step absorption process is possible*

4.3   Solar Radiation

Sunlight is caused by blackbody radiation from the outer layer of the sun. At a temperature of approximately 5250°C, this layer emits a spectrum as shown in Figure 4.6, which represents the solar spectrum in space and is relevant to solar cells used on satellites and space stations. Terrestrial solar cells, however, rely on the terrestrial solar spectrum, which suffers substantial attenuation at certain wavelengths. In particular, water molecules absorb strongly in four infrared bands between 800 and 2000 nm as shown.

4.4   Solar Cell Design and Analysis

The design of a practical silicon solar cell can now be considered. In order for light to reach the junction area of the p-n junction, the junction should be close to the surface of the semiconductor. The junction area must be large enough to capture the desired radiation. This dictates a thin n- or p-region on the illuminated side of the solar cell. A significant challenge is to enable the thin region to be sufficiently uniform in potential to allow the junction to function over its entire area. If a contact material is applied to the surface of the cell, sunlight will be partly absorbed in the contact material. The common solution to this is to make the thin region of the silicon as conductive as possible by doping it heavily. In this way, the highly doped thin region simultaneously serves as a front electrode with high lateral conductivity (conductivity in a plane parallel to the plane of the junction) *and* as one side of the p-n junction. Since n-type silicon has higher electron mobility and therefore
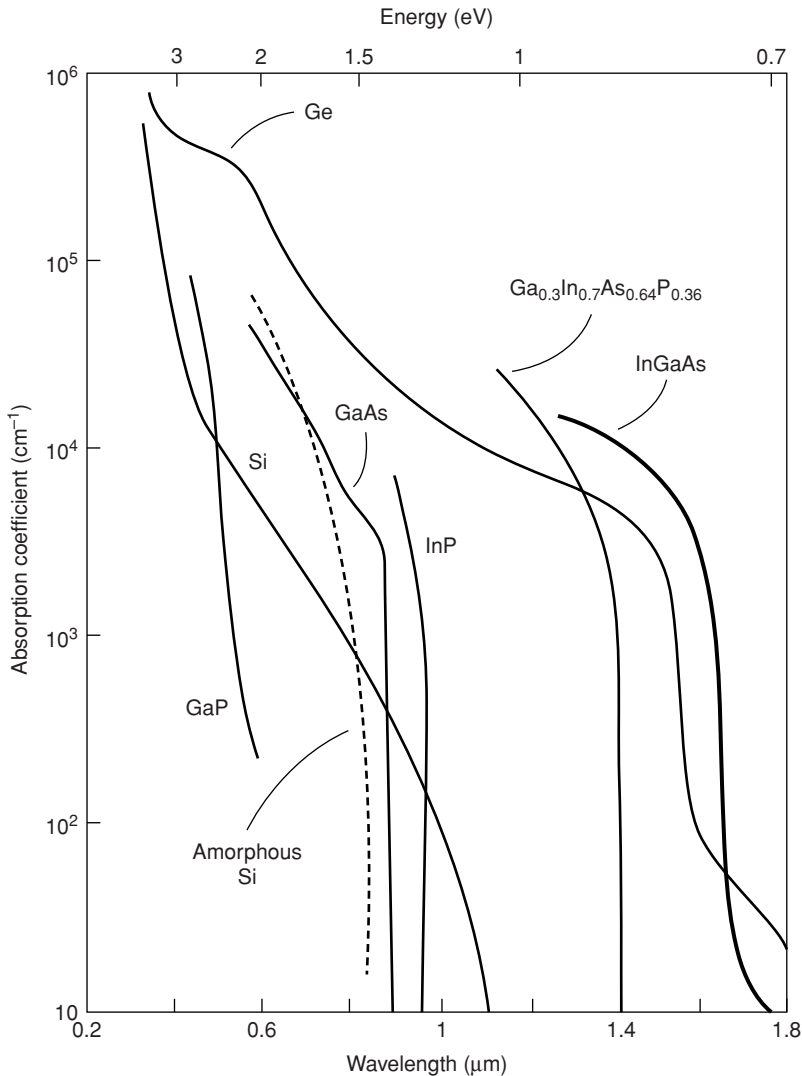
Figure 4.5   *Absorption coefficients covering the solar spectral range for a range of semiconductors. Note the absorption tails in silicon and germanium arising from two-step absorption processes. Amorphous silicon is a non-crystalline thin film that has different electron states and hence different absorption coefficients compared with single-crystal silicon. Reprinted from Shur, M., Physics of semiconductor devices. Copyright (1990) with permission from Prentice Hall, USA*

higher conductivity than is achieved by the lower mobility of holes in p-type material, the thin top layer in silicon solar cells is preferably n-type in practice.

A crystalline silicon solar cell is shown in Figure 4.7. It consists of a thin n^+ front layer. A metal grid is deposited on this layer and forms an ohmic contact to the n^+ material. The areas on the n^+ front layer that are exposed to sunlight are coated with an antireflection
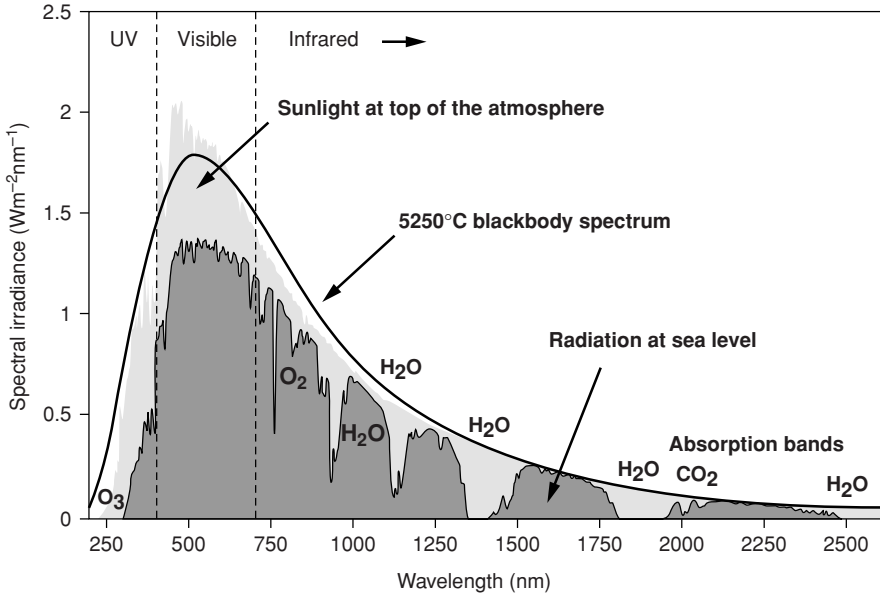
Figure 4.6   *Solar radiation spectrum for a 5250°C blackbody, which approximates the space spectrum of the sun, as well as a spectrum at the earth's surface that survives the absorption of molecules such as H_2O and CO_2 in the earth's atmosphere. Note also the substantial ozone (O_3) absorption in the UV part of the spectrum. Reproduced from www.global warmingart.com/wiki/File:Solar_Spectrum_png. Copyright (2011) globalwarmingart.com*
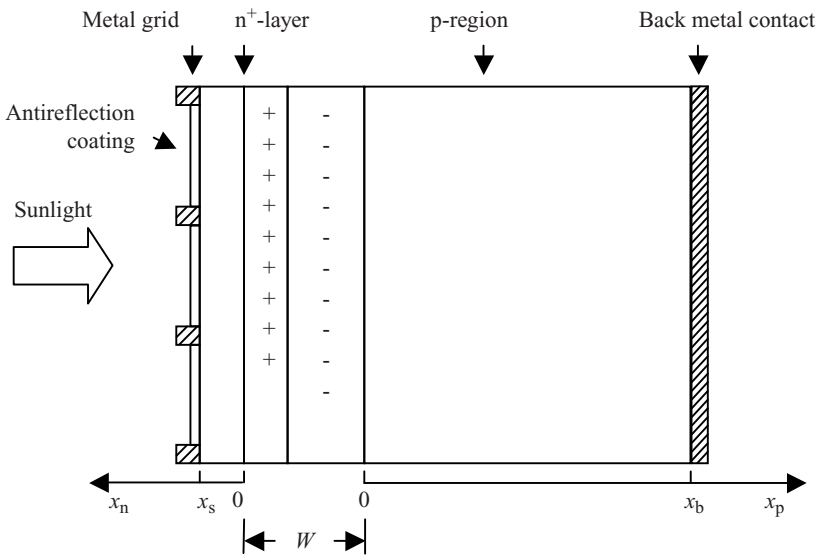
Figure 4.7   *Cross-section of a silicon solar cell showing the front contact metal grid that forms an ohmic contact to the n^+-layer. The depletion region at the junction has width W*

layer. The simplest such layer is a quarter wavelength in thickness such that incident light waves reflecting off the front and back surfaces of this layer can substantially cancel each other (see Problem 4.3). The metal grid does block some sunlight; however, in practice the grid lines are narrow enough to prevent excessive light loss. A thick p-type region absorbs virtually all the remaining sunlight, and is contacted by a rear metal ohmic contact.

Because most of the photons are absorbed in the thick p-type layer, most of the minority carriers that need to be collected are electrons. The goal is to have these electrons reach the front contact. There will, however, also be some minority holes generated in the n$^+$ region that ideally reach the p-region.

Sunlight entering the solar cell will be absorbed according to the relationship introduced in Section 1.12

$$I(x) = I_0 e^{-\alpha x}$$

In order to simplify the treatment of the solar cell we will assume that the optical generation rate G is uniform throughout the p-n junction. This implies that the absorption constant α is small. Real solar cells are approximately consistent with this assumption for photons of longer wavelengths of sunlight very close in energy to the bandgap. Shorter wavelengths, however, should really be modelled as a rapidly decaying generation rate with depth.

We will also start by assuming that the relevant diffusion lengths of minority carriers in both the n-type and p-type regions are much shorter than the thicknesses of these regions. This means that the p-n junction may be regarded as possessing semi-infinite thickness as far as excess minority carrier distributions are concerned, and in Figure 4.7 the front surface and back surface at $x_n = x_s$ and $x_p = x_b$ respectively are far away from regions containing excess carriers.

For the n-side, the diffusion equation (Equation 1.64a) for holes may be rewritten as

$$D_p \frac{d^2 \delta p(x_n)}{dx^2} = \frac{\delta p(x_n)}{\tau_p} - G \tag{4.1}$$

The term G must be subtracted from the hole recombination rate $\dfrac{\delta p(x_n)}{\tau_p}$ because it is the additional hole generation rate. The solution to this equation where

$$L_p = \sqrt{D_p \tau_p}$$

is

$$\delta p(x_n) = G\tau_p + C \exp\left(\frac{x}{L_p}\right) + D \exp\left(\frac{-x_n}{L_p}\right) \tag{4.2}$$

Note that the solution is the same as Equation 1.65a except for the added term $G\tau_p$. See Problem 4.4.

The boundary conditions we shall satisfy are:

$$\delta p(0) = p_n \left(\exp \frac{qV}{kT} - 1 \right)$$

and

$$\delta p(x_n \to \infty) = G\tau_p$$
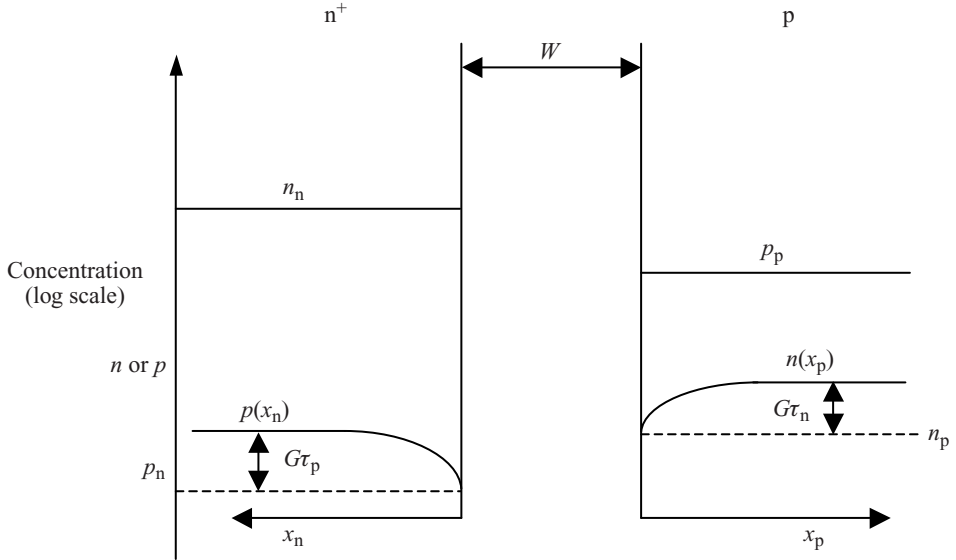
Figure 4.8   *Concentrations are plotted on a log scale to allow details of the minority carrier concentrations as well as the majority carrier concentrations to be shown on the same plot. Note that the n+-side has higher majority carrier concentration and lower equilibrium minority carrier concentration than the more lightly doped p-side corresponding to Figure 4.7. Diffusion lengths are assumed to be much smaller than device dimensions. The p-n junction is shown in a short circuit condition with V = 0*

The first boundary condition is as discussed in Section 2.5. The dynamic equilibrium in the depletion region still determines the carrier concentrations at the depletion region boundaries. For the second boundary condition $G\tau_p$ is the excess carrier concentration optically generated far away from the depletion region (see Equation 1.51).

Substituting these two boundary conditions into Equation 4.2 we obtain (see problem 4.4)

$$\delta p(x_n) = G\tau_p + \left[p_n \left(\exp\left(\frac{qV}{kT} \right) - 1 \right) - G\tau_p \right] \exp\left(-\frac{x_n}{L_p} \right) \tag{4.3a}$$

The analogous equation for the p-side is

$$\delta n(x_p) = G\tau_n + \left[n_p \left(\exp\left(\frac{qV}{kT} \right) - 1 \right) - G\tau_n \right] \exp\left(-\frac{x_p}{L_n} \right) \tag{4.3b}$$

Note that if $V = 0$, Equations 4.3a and 4.3b yield $\delta p(x_n) = G\tau_p$ and $\delta n(x_p) = G\tau_n$ for large values of x_n and x_p respectively. In addition at $V = 0$ these equations give zero for both $\delta p(x_n = 0)$ and $\delta n(x_p = 0)$. We can show this more clearly in Figure 4.8 for an illuminated p-n junction under short circuit conditions ($V = 0$). We use Equation 4.3 to plot $p(x_n) = \delta p(x_n) + p_n$ and $n(x_p) = \delta n(x_p) + n_p$, where p_n and n_p are the equilibrium minority carrier concentrations.

Having determined the minority carrier concentrations we can now determine diffusion currents $I_n(x)$ and $I_p(x)$ using the equations for diffusion currents (Equation 1.54). By substitution of Equations 4.3a and 4.3b into Equation 1.54 we obtain

$$I_p(x_n) = \frac{qAD_p}{L_p}p_n\left(\exp\left(\frac{qV}{kT}\right) - 1\right)\exp\left(-\frac{x_n}{L_p}\right) - qAGL_p\exp\left(\frac{x_n}{L_p}\right) \quad (4.4a)$$

for holes diffusing in the n-side and

$$I_n(x_p) = \frac{qAD_n}{L_n}n_p\left(\exp\left(\frac{qV}{kT}\right) - 1\right)\exp\left(-\frac{x_p}{L_n}\right) - qAGL_n\exp\left(\frac{x_p}{L_n}\right) \quad (4.4b)$$

for electrons diffusing in the p-side (see Problem 4.5). Note that the first terms in these equations are identical with Equations 2.21d and 2.21c for an unilluminated diode.

Since there is uniform illumination, we need also to consider generation in the depletion region. We shall neglect recombination of electron-hole pairs since W is much smaller than the carrier diffusion lengths. This means that every electron and every hole created in the depletion region contributes to diode current. The generation rate G must be multiplied by depletion region volume WA to obtain the total *number of carriers generated per unit time* in the depletion region. Carrier current optically generated from inside the depletion region therefore becomes the total charge generated per unit time or

$$I_{(depletion)} = qGWA \quad (4.5)$$

Although both an electron and a hole are generated by each absorbed photon, each charge pair is only counted once: one generated electron drifts to the n-side metal contacts, flows through the external circuit, and returns to the p-side. In the meantime, one hole drifts to the p-side metal contact and is available there to recombine with the returning electron.

It therefore follows using Equation 2.22 that the total diode current becomes

$$I = I_0\left(\exp\left(\frac{qV}{kT}\right) - 1\right) - I_L \quad (4.6)$$

where I_L, the current optically generated by sunlight, has three components, from the n-side, from the depletion region, and from the p-side respectively. Using Equation 4.5 as well as the second terms from Equations 4.4a and 4.4b at $x_n = x_p = 0$, we obtain

$$I_L = qAG(L_n + W + L_p) \quad (4.7)$$

which confirms that Figure 4.2 is valid and the $I-V$ characteristic is shifted vertically (by amount I_L) upon illumination. Of the three terms in Equation 4.7, the second term is generally smallest due to small values of W compared to carrier diffusion lengths, and since electron mobility and diffusivity values are higher than for holes the first term will be larger than the third term. It is reasonable that diffusion lengths L_n and L_p appear in Equation 4.7: carriers must cross over the depletion region to contribute to solar cell output current. They have an opportunity to diffuse to the depletion region before they drift across it, and the diffusion lengths are the appropriate length scales over which this is likely to occur.
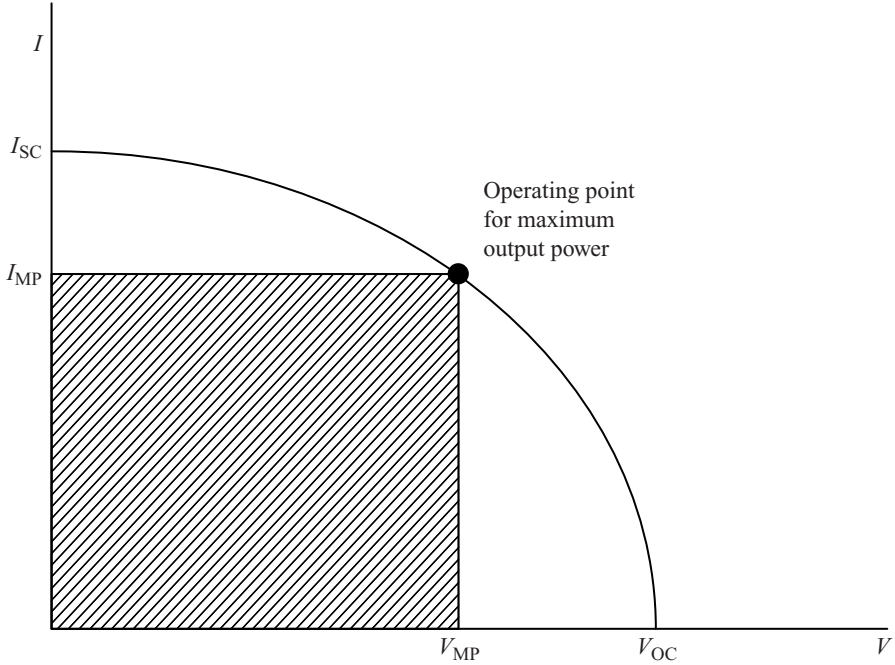
Figure 4.9   *Operating point of a solar cell. The fourth quadrant in Figure 4.2 is redrawn as a first quadrant for convenience. Open circuit voltage V_{OC} and short circuit current I_{OC} as well as current I_{MP} and voltage V_{MP} for maximum power are shown. Maximum power is obtained when the area of the shaded rectangle is maximized*

The solar cell *short circuit current* I_{SC} can now be seen to be the same as I_L by setting $V = 0$ in Equation 4.6. Hence

$$I_{SC} = qAG(L_n + W + L_p) \tag{4.8}$$

In addition, the solar cell *open circuit voltage* V_{OC} can be found by setting $I = 0$ in Equation 4.6 and solving for V to obtain

$$V_{OC} = \frac{kT}{q} \ln\left(\frac{I_L}{I_0} + 1\right) \tag{4.9}$$

These quantities are plotted in Figure 4.9 together with the solar cell operating point.

The fill factor *FF* is defined as

$$FF = \frac{I_{MP} V_{MP}}{I_{SC} V_{OC}} \tag{4.10}$$

In crystalline silicon solar cells *FF* may be in the range of 0.7 to 0.85.

Example 4.1

An abrupt silicon p-n junction solar cell at room temperature is exposed to sunlight. Assume that the sunlight is uniformly intense throughout the silicon yielding an optical generation rate of 5×10^{19} cm^{-3} s^{-1}. The solar cell has a junction area of 100 cm^2, a depletion region width of 3 μm and reverse saturation current density of $J_0 = 1 \times 10^{-11}$A cm^{-2}. The silicon has a carrier lifetime of 2×10^{-6} s.

(a) Find the optically generated current that is generated inside the depletion region.
(b) Find the total optically generated current.
(c) Find the short-circuit current.
(d) Find the open-circuit voltage.
(e) If the solar cell fill factor is 0.75, find the maximum power available and discuss this in terms of the expected total available sunlight.

Solution

(a)

$$I_{(depletion)} = qGWA = 1.6 \times 10^{-19} \text{ C} \times 5 \times 10^{19} \text{ cm}^{-3} \text{ s}^{-1} \times 3 \times 10^{-4} \text{ cm}$$
$$\times 100 \text{ cm}^2 = 0.24 \text{ A}$$

(b) From Example 2.3,

$$L_n = \sqrt{D_n \tau_n} = \sqrt{3.51 \times 10^1 \text{ cm}^2 \text{ s}^{-1} \times 2 \times 10^{-6} \text{ s}} = 8.38 \times 10^{-3} \text{ cm}$$

and

$$L_p = \sqrt{D_p \tau_p} = \sqrt{1.25 \times 10^1 \text{ cm}^2 \text{ s}^{-1} \times 2 \times 10^{-6} \text{ s}} = 5.00 \times 10^{-3} \text{ cm}$$

and therefore

$$I_L = qAG(L_n + W + L_p)$$

$$= 1.6 \times 10^{-19} \text{ C} \times 100 \text{ cm}^2 \times 5 \times 10^{19} \text{ cm}^{-3} \text{ s}^{-1}(8.38 \times 10^{-3} + 3 \times 10^{-4}$$
$$+ 5.00 \times 10^{-3} \text{ cm}) = 10.9 \text{ A}$$

(c) The short circuit current I_{SC} is the same as I_L. Therefore $I_{SC} = 10.9$ A.
(d) Open-circuit voltage:

$$I_0 = J_0 A = 1 \times 10^{-11}A \text{ cm}^{-2} \times 100 \text{ cm}^2 = 1 \times 10^{-9} \text{ A}$$

and now

$$V_{OC} = \frac{kT}{q} \ln\left(\frac{I_L}{I_0} + 1\right) = 0.026 \text{ V} \ln\left(\frac{10.9 \text{ A}}{1 \times 10^{-9}A} + 1\right) = 0.601 \text{ V}$$

(e) The maximum output power is obtained at the operating point V_{MP} and I_{MP}. Therefore

$$I_{MP} V_{MP} = FF \times I_{SC} V_{OC} = 0.75 \times 10.9 \text{ A} \times 0.601 \text{ V} = 4.91 \text{ W}$$

The available sunlight per square metre for full sun conditions on the earth's surface is approximately 1000 W, which yields 10 W over an area of 100 cm^2. The best silicon solar cell, however, is not more than 25% efficient and therefore the most electrical power that we could expect to be available from the solar cell should be closer to 2.5 W. The model we have used assumes a uniform optical generation rate inside the silicon. This is not realistic since sunlight will be absorbed and a decreasing optical generation rate with depth will exist in reality, which will decrease the available power.

In practice achieving an electrical output power of 4.91 W from a silicon solar cell of 100 cm^2 would require a concentration of sunlight by a factor of approximately two using reflective or refractive optical concentrators. See Section 4.14.

4.5   Thin Solar Cells

Since practical solar cells have a very thin n$^+$ layer, as shown in Figure 4.7, and even the p-side may be small in thickness, we will now take this into consideration in the device model. We will remove the restriction that the relevant diffusion lengths of minority carriers in both the n-type and p-type regions are much shorter than the thicknesses of these regions. This means that in Figure 4.7 the front surface and back surface at $x_n = x_s$ and $x_p = x_b$ respectively, must be assigned suitable boundary conditions in order to calculate minority carrier concentrations.

We have discussed semiconductor surfaces in Section 1.19. We can apply Equation 1.47 to the front interface at $x_n = x_s$ and we can write

$$-D_n \frac{d\delta p(x_s)}{dx} = S_f \delta p(x_s) \tag{4.11}$$

where S_f is the *effective front surface recombination velocity*, which is a modified value of surface recombination velocity because the front surface is not a free surface and surface states will be influenced by the antireflection layer as well as the small areas occupied by the ohmic contacts. The minority carriers in the n$^+$-region are holes.

The back metal interface must also be modelled. As with the front surface, we will use an *effective back surface recombination velocity* S_b to describe this and hence at the back surface,

$$-D_n \frac{d\delta n(x_b)}{dx} = S_b \delta n(x_b) \tag{4.12}$$

where the minority carriers in the p-region are electrons.

Consider the p-type material as shown in Figure 4.7. At $x_p = 0$ the excess minority carrier concentration in an unilluminated junction at voltage V is given from Equation 2.17b as

$$\delta n_{p(x_p=0)} = n_p \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

and at $x_p = x_b$, δn_p will depend on the value of S_b. If we examine the case in which S_b is very large then from Equation 4.12 it follows that

$$\delta n_{p(x_p = x_b)} \cong 0$$

These two boundary conditions may be substituted into the general solution of the diffusion equation for electrons. In Equation 1.65a we have written this for holes. For electrons it becomes

$$\delta n_p(x_p) = A \exp\left(\frac{-x_p}{L_n}\right) + B \exp\left(\frac{x_p}{L_n}\right) \tag{4.13}$$

where $L_n = \sqrt{D_n \tau_n}$. We now need to consider both terms since the length of the p-type material is finite and we are not justified in assuming that $B = 0$. The two boundary conditions give us two equations

$$n_p \left(\exp\left(\frac{qV}{kT}\right) - 1\right) = A + B \tag{4.14}$$

and

$$0 = A \exp\left(\frac{-x_b}{L_n}\right) + B \exp\left(\frac{x_b}{L_n}\right) \tag{4.15}$$

Multiplying Equation 4.14 by $\exp\left(\dfrac{x_b}{L_n}\right)$ and subtracting it from Equation 4.15 we can solve for A and obtain

$$A = \frac{\exp\left(\dfrac{x_b}{L_n}\right)}{\exp\left(\dfrac{x_b}{L_n}\right) - \exp\left(-\dfrac{x_b}{L_n}\right)} n_p \left(\exp\left(\frac{qV}{kT}\right) - 1\right)$$

By multiplying Equation 4.14 by $\exp\left(-\dfrac{x_b}{L_n}\right)$ we can similarly solve for B and obtain

$$B = \frac{\exp\left(-\dfrac{x_b}{L_n}\right)}{\exp\left(-\dfrac{x_b}{L_n}\right) - \exp\left(\dfrac{x_b}{L_n}\right)} n_p \left(\exp\left(\frac{qV}{kT}\right) - 1\right)$$

Substituting A and B into Equation 4.13 we obtain

$$\delta n_p(x_p) = \frac{n_p \left(\exp\left(\dfrac{qV}{kT}\right) - 1\right)}{\exp\left(\dfrac{x_b}{L_n}\right) - \exp\left(-\dfrac{x_b}{L_n}\right)} \left[\exp\left(\frac{x_b}{L_n}\right)\exp\left(-\frac{x_p}{L_n}\right) - \exp\left(-\frac{x_b}{L_n}\right)\exp\left(\frac{x_p}{L_n}\right)\right]$$

$$\tag{4.16}$$

The identical method may be used to find the minority hole concentration in the n$^+$ material. We will assume that S_f is very large, which implies that

$$\delta p_{n(x_n = x_s)} \cong 0$$
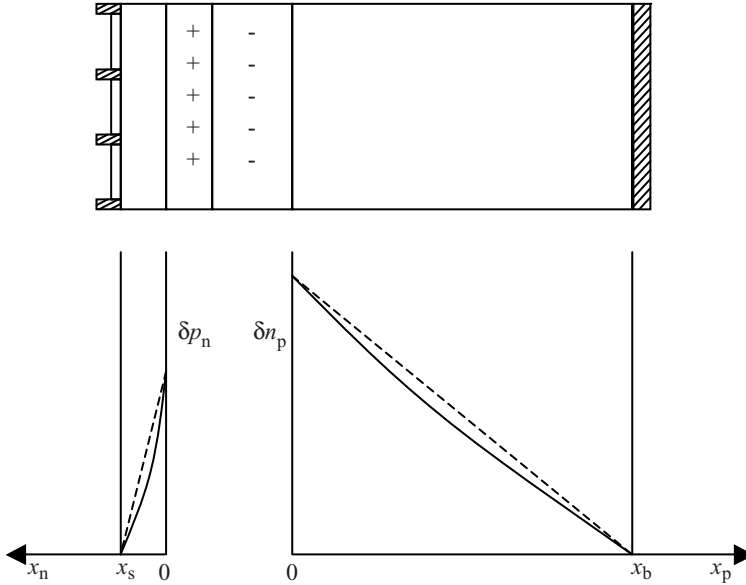
Figure 4.10   *Excess minority carrier concentrations for a solar cell having dimensions* x_s *and* x_b *that are small compared to the carrier diffusion lengths. Very high values of surface recombination velocity are present. Note that a forward bias is assumed, and there is no illumination*

and we obtain

$$\delta p_n(x_n) = \frac{p_n\left(\exp\left(\dfrac{qV}{kT}\right) - 1\right)}{\exp\left(\dfrac{x_s}{L_p}\right) - \exp\left(-\dfrac{x_s}{L_p}\right)}\left[\exp\left(\frac{x_s}{L_p}\right)\exp\left(-\frac{x_n}{L_p}\right) - \exp\left(-\frac{x_s}{L_p}\right)\exp\left(\frac{x_n}{L_p}\right)\right]$$

(4.17)

Equations 4.16 and 4.17 become meaningful if they are plotted. If we substitute $x_p = x_b$ into Equation 4.16 we obtain $\delta n_p(x_b) = 0$ as expected, and if we substitute $x_n = x_s$ into Equation 4.17 we obtain $\delta p_n(x_n) = 0$ as expected. The resulting plots are shown in Figure 4.10 for the p-n junction in a forward bias condition without illumination. If x_b and x_s are much smaller than L_n and L_p respectively, then $\delta n_p(x_p)$ and $\delta p_n(x_n)$ become almost straight lines as shown. This can be understood from Equation 4.13 since exponential functions are approximately straight lines for small arguments. (The function $\exp(x)$ may be approximated as $1 + x$ for small values of x.)

It is clear that the carrier concentrations of both minority electrons and minority holes decrease rapidly as we move away from either side of the depletion region. This means that minority diffusion currents are enabled that flow *away* from the junction. The thinner the solar cell becomes, the steeper these decreases become. Minority currents flow in the opposite directions to the directions that we require: in a solar cell the minority currents should flow *towards* the junction.

If illumination were incident uniformly throughout the solar cell in Figure 4.10, excess carriers would be generated. If the solar cell was operated with an electrical load, the

solar cell voltage would decrease and the minority carrier concentrations near the depletion region edges would decrease. The result is that upon illumination of the solar cell of Figure 4.10, surface recombination would compete strongly with the desired carrier drift across the depletion region and solar cell performance would be poor.

For this reason, thin solar cells having high values of S_B and S_F are not ideal. We will now examine a second limiting case in which S_B and S_F are assumed to be zero. If we once again look at the p-side of the junction, minority current flow to the back surface at $x_B = 0$ is zero, and from Equation 4.12

$$-D_n \frac{d\delta n(x_b)}{dx} = 0$$

This implies that the excess minority carrier concentration δn is independent of x_b. $\delta n(x_b)$ is therefore a constant and is equal to the value of δn at the edge of the depletion region. From Equation 2.17b we have in the p-side

$$\delta n_p(x_p) = n_p \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

and in the n-side

$$\delta p_n(x_n) = p_n \left(\exp\left(\frac{qV}{kT}\right) - 1 \right)$$

These straight lines are shown in Figure 4.11 for a solar cell without illumination in a forward bias condition. Note that if the solar cell were connected to an electrical load
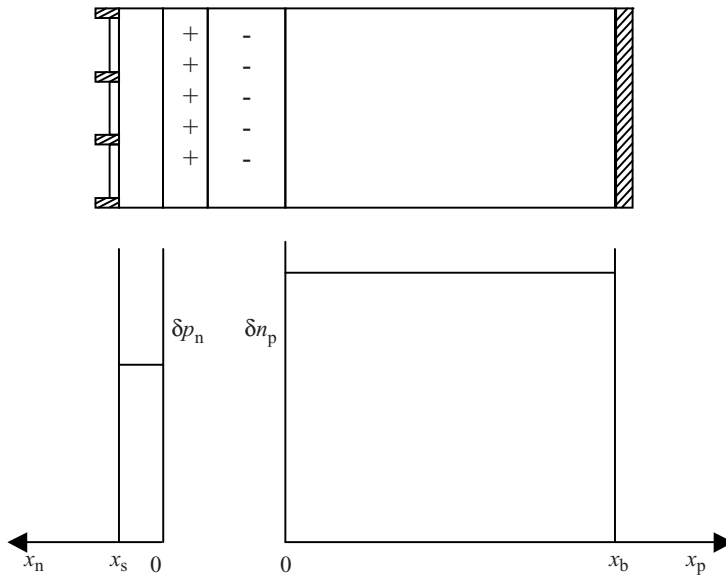
Figure 4.11 *Excess minority carrier concentrations for a solar cell having dimensions* x_s *and* x_b *that are small compared to the carrier diffusion lengths. Zero surface recombination velocity is assumed, which means that there is no drop-off of carrier concentrations towards the front and back surfaces*

and illuminated, the minority carrier concentrations would slope downwards towards the depletion region to approach (but not reach) the condition at $V = 0$ of Figure 4.8, and the generated minority carriers would have an opportunity to diffuse towards the junction without competition from surface recombination. Although some bulk recombination is always present, the reduced thickness of the solar cell would reduce recombination in general. The thickness of the solar cell could now be set to allow adequate optical absorption of the sunlight without concern for surface recombination. In practice effective surface recombination velocities must be minimized. This will be further discussed in Section 4.6.

4.6   Solar Cell Generation as a Function of Depth

The efficiency of electron and hole pair collection in a solar cell can be analysed as a function of the depth at which carriers are generated by an absorbed photon. We can assume that generation occurs *at only one depth*, and then determine the probability that carriers are collected and cross over the junction. We will simplify the problem by assuming that

(a) The solar cell is operating under short circuit conditions.
(b) The diffusion length is much smaller than the solar cell thickness.

Since most carriers are absorbed in the p-type layer we will confine our attention to the p-side of the junction, although the analysis may readily be extended to the n-side and to the depletion region. The assumed generation rate will resemble a delta function, as shown in Figure 4.12. Note that although this situation is physically not realizable it does serve to illustrate how the ability to collect minority carriers depends on where they are generated.
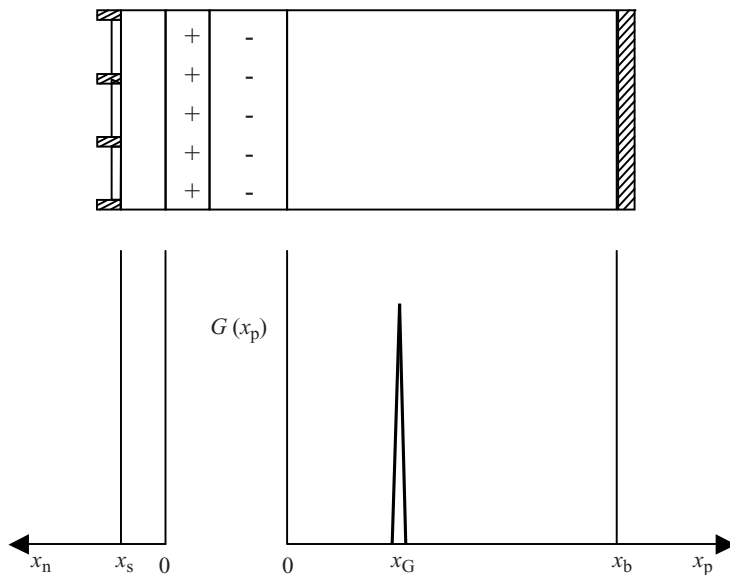
Figure 4.12   *Generation rate as a function of depth showing zero generation except at a specific depth x_G*

For $0 \leq x_p \leq x_G$ we can determine the carrier concentration using the diffusion equation and obtain the general solution of Equation 4.13:

$$\delta n_p(x_p) = \alpha \exp\left(\frac{x_p}{L_n}\right) + \beta \exp\left(\frac{-x_p}{L_n}\right)$$

Under short circuit conditions, excess carrier concentration at $x_p = 0$ will be zero and hence

$$\beta = -\alpha$$

Now,

$$\delta n_p(x_p) = \alpha \left\{\exp\left(\frac{x_p}{L_n}\right) - \exp\left(\frac{-x_p}{L_n}\right)\right\} \tag{4.18}$$

For $x_G \leq x_p \leq x_b$ the general solution to the diffusion equation will be

$$\delta n_p(x_p) = \gamma \exp\left(\frac{-x_p}{L_n}\right) \tag{4.19}$$

because the function must fall to zero for large values of x_p. To be a continuous function we require that Equations 4.18 and 4.19 are equal at $x_p = x_G$ and

$$\alpha \left\{\exp\left(\frac{x_G}{L_n}\right) - \exp\left(\frac{-x_G}{L_n}\right)\right\} = \gamma \exp\left(\frac{-x_G}{L_n}\right)$$

or

$$\gamma = \alpha \left\{\exp\left(\frac{2x_G}{L_n}\right) - 1\right\}$$

which may be used to replace γ in Equation 4.19 resulting, for $x_G \leq x_p \leq x_b$, in

$$\delta n_p(x_p) = \alpha \left\{\exp\left(\frac{2x_G}{L_n}\right) - 1\right\} \exp\left(\frac{-x_p}{L_n}\right) \tag{4.20}$$

The functions $\delta n_p(x_p)$ from Equations 4.18 and 4.20 are plotted in Figure 4.13.

Now the diffusion current density arising from these carrier concentrations may be calculated in the normal manner using

$$J_n = q D_n \frac{dn_p}{dx_P}$$

At $x_p = 0$ the diffusion current is determined using the derivative of Equation 4.18, and we obtain

$$\left|J_n(x_p = 0)\right| = 2\alpha q \frac{D_n}{L_n} \tag{4.21}$$

At $x_p = x_G$ there are two components of the diffusion current, both flowing towards the generation zone at $x_p = x_G$. (The fluxes of excess electrons actually flow *away* from x_G in the directions of decreasing carrier concentration; however, the currents flow in directions opposite to the fluxes.) The *magnitudes* of these currents are shown in Figure 4.13 and are determined using the slopes at $x_p = x_G$ of Equations 4.18 and 4.20, respectively.
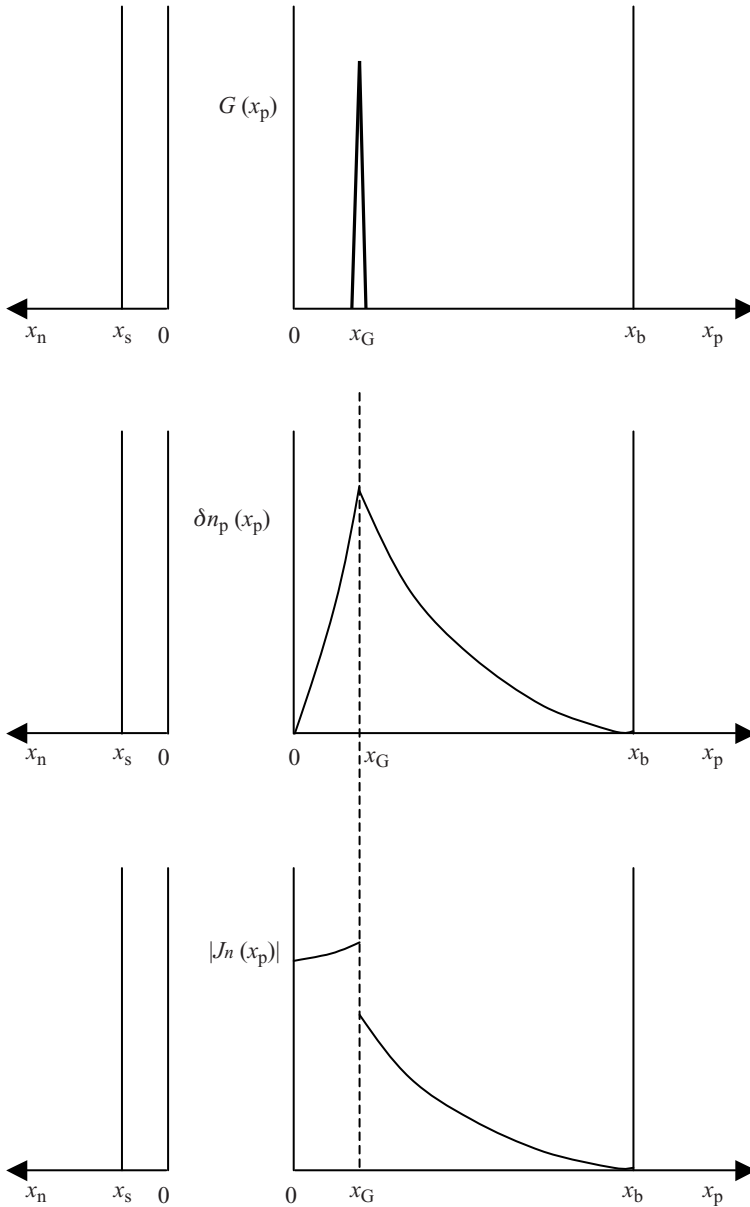
Figure 4.13   *Generation rate, excess carrier concentration and magnitude of the diffusion current density as a function of position in the p-side of the solar cell. Note that the diffusion current is positive for $0 \leq x_p \leq x_G$ and negative for $x_G \leq x_p \leq x_b$*

We obtain a current magnitude due to the electron flux flowing to the left of x_G, using Equation 4.18, of

$$\left| J_{1n}(x_p = x_G) \right| = q\alpha \frac{D_n}{L_n} \left\{ \exp\left(\frac{x_G}{L_n} \right) + \exp\left(\frac{-x_G}{L_n} \right) \right\}$$

and current magnitude due to electron flux flowing to the right of x_G, using Equation 4.20, of

$$\left| J_{2n}(x_p = x_G) \right| = q\alpha \frac{D_n}{L_n} \left\{ \exp\left(\frac{x_G}{L_n} \right) - \exp\left(\frac{-x_G}{L_n} \right) \right\}$$

The total generated current density at x_b is the sum of these two current magnitudes, or

$$|J_{total}(x_b)| = 2q\alpha \frac{D_n}{L_n} \exp\left(\frac{x_G}{L_n} \right) \tag{4.22}$$

The fraction η of this total current that reaches the depletion region is obtained by dividing Equation 4.21 by Equation 4.22 to yield

$$\eta = \exp\left(-\frac{x_G}{L_n} \right)$$

This means that the contribution to the usable current flow decreases exponentially as a function of the distance between the depletion region and the position of electron-hole pair (EHP) generation in the p-type region. For optimum performance in solar cells of semi-infinite thickness ($x_b \gg L_n$) we therefore require that the absorption depth of photons is smaller than the diffusion length L_n.

In silicon, photons are absorbed over a depth of approximately 100 μm and diffusion lengths in the range of 1 mm are achievable. This implies that the thickness of the solar cell should be more than 1 mm. Since silicon wafers under 0.2 mm in thickness may be cut and processed into solar cells in large volumes it is clear that a low effective surface recombination velocity at the rear contact is required. This may be achieved by forming a p$^+$ doped region near the back contact. Figure 4.14 shows the resulting solar cell band diagram. A potential step near the back contact is formed that helps to prevent minority electrons from reaching the back silicon-metal interface due to the *back surface electric field* that is created at the step. Much lower effective rear surface recombination velocities result from this and the rear p$^+$-region is therefore part of the standard solar cell design.

4.7   Solar Cell Efficiency

There are some fundamental constraints that set an upper limit on the efficiency that can be achieved from a single-junction solar cell.

The open circuit voltage V_{OC} must be as large as possible for maximum efficiency. From Equation 4.9 this means that I_0 should be as small as possible. We can rewrite I_0 from Equation 2.22, and using $n_p = \dfrac{n_i^2}{N_A}$ and $p_n = \dfrac{n_i^2}{N_D}$ we obtain

$$I_0 = qA \left(\frac{D_n}{L_n} \frac{n_i^2}{N_A} + \frac{D_p}{L_p} \frac{n_i^2}{N_D} \right) \tag{4.23}$$
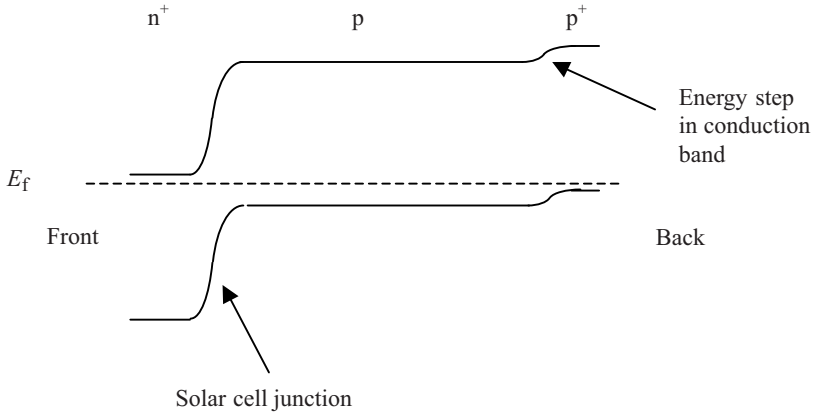
Figure 4.14   *Back surface field formed by a p$^+$ doped region near the back of the solar cell. A potential energy step that generates a built-in electric field decreases the likelihood of electrons reaching the back surface of the silicon*
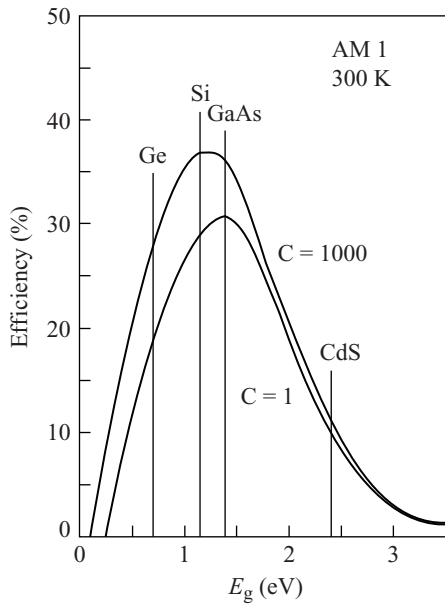
Figure 4.15   *Efficiency limit of solar cells based on a number of well-known semiconductors. Note the increase in efficiency potentially available if the sunlight intensity is increased to 1000 times the normal sun intensity. This increase can be understood since from (4-8) I_{SC} is proportional to G, and in addition from (4-7) and (4-9)V_{OC} also increases gradually with G. Hence $I_{SC}V_{OC}$ increases more quickly than the solar radiation intensity and the cell efficiency will increase. Reprinted from Sze, S.M., Semiconductor Devices: Physics and Technology, 9780471837046. Copyright (1985) with permission from John Wiley & Sons, New York*

which clearly shows the sensitivity of I_0 on n_i. From Equation 1.37, n_i decreases exponentially as energy gap E_g increases and as T decreases:

$$n_i = N_c N_v \exp\left(\frac{-E_g}{2kT}\right) \tag{4.24}$$

The other quantity that must be as large as possible is the short circuit current I_{SC}, which depends linearly on the optical generation rate G from Equation 4.8. G will decrease once E_g exceeds the photon energy of incoming photons. This means that there is an optimal energy gap, as shown in Figure 4.15, for a number of semiconductors.

There are two major contributions to efficiency loss in p-n junction solar cells. The first arises from photons having energies higher than E_g. The extra photon energy becomes carrier kinetic energy, which quickly gets converted to heat as the carriers relax or *thermalize* to their lower energy states. This happens before the carriers can be collected and utilized. This would not be an issue if the solar spectrum was monochromatic, but the broad blackbody solar spectrum immediately limits efficiency values to under 50% for a single-junction solar cell.

The second is the discrepancy between the optimum operating voltage V_{MP} and the bandgap E_g of the semiconductor used. A minimum photon energy corresponding to E_g is needed to create electron-hole pairs; however, V_{MP} is less than E_g. For example, in silicon, $\frac{V_{MP}}{E_g} \cong \frac{0.6}{1.1}$, or about 55%. This means that the overall efficiency of a single-junction solar cell will be under 30%. Silicon solar cells have reached approximately 25% efficiency in the lab, and 24% in production. See Problem 4.14. GaAs-based solar cells having an optimal direct gap have reached only slightly higher values of 26%.

Figure 4.16 shows the trend towards higher efficiencies in solar cell technology as time progresses. Note that the higher efficiency solar cell types are *multi-junction* solar cells, which will be covered in Section 4.13. Production single-crystal silicon solar cells are in the range of 16–24% efficient.

Solar cell efficiency also depends on the operating temperature of the solar cell p-n junction. In general there will be a decrease in efficiency at higher temperatures due to the increase in I_0. This means that it may be advantageous to generate solar power at higher latitudes where ambient temperatures are lower. The disadvantage, however, is the lower angle of the sun relative to locations near the Equator. See Example 4.2.

Example 4.2

For the solar cell of Example 4.1,

(a) Find the saturation currents at $-50°C$ and at $+100°C$ relative to the saturation current at room temperature (300 K).
(b) Find the predicted maximum output power at $-50°C$ and at $+100°C$ and compare to the maximum output power at room temperature (300 K).

Make and state any necessary assumptions

Solution

(a) In order to obtain saturation current I_0 at various temperatures we will assume that N_c, N_v, E_g, D_n, D_p, L_n and L_p are temperature independent. This is not
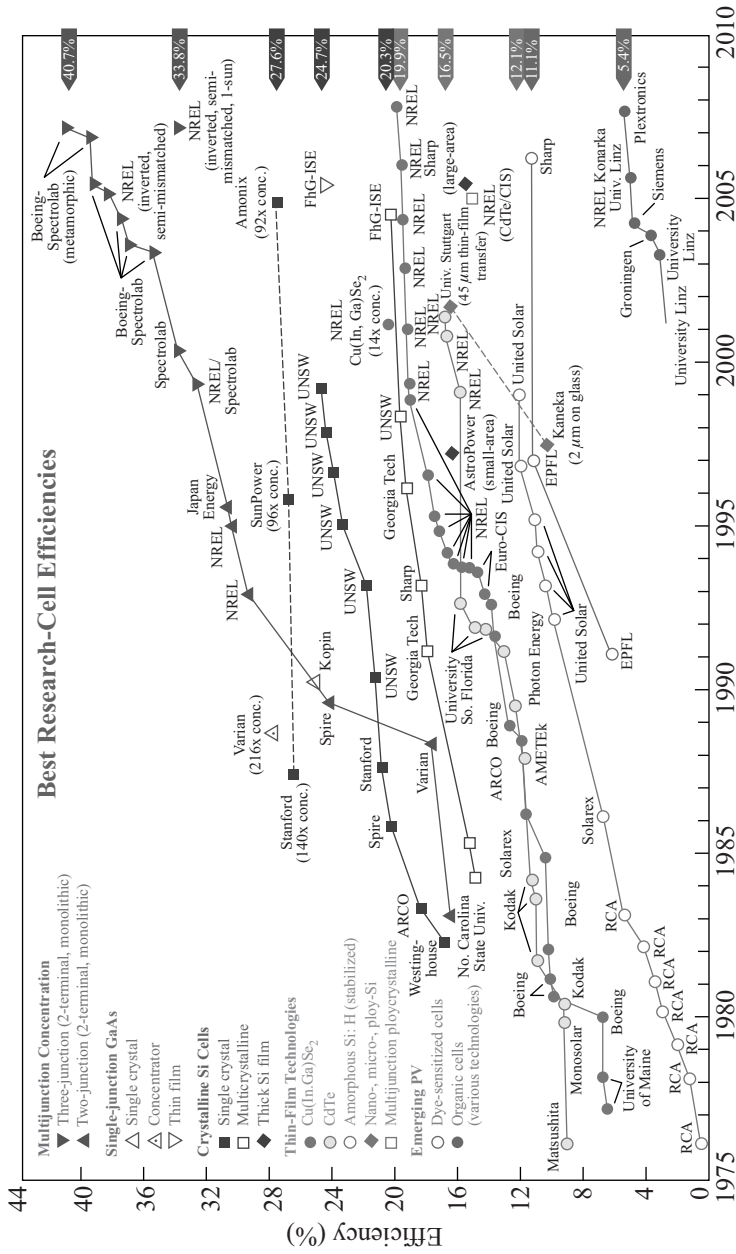
Figure 4.16 *Best research solar cell efficiencies achieved in a given year. Courtesy of United States Department of Energy, Property of the U.S Federal Government. Data compiled by Lawrence Kazmerski, National Renewable Energy Laboratory (NREL)*

strictly true since E_g decreases with increasing temperature as lattice vibrations decrease the average periodic potential in the semiconductor crystal. In addition there are changes in diffusion length as carrier lifetime and carrier diffusivity are temperature dependent. In spite of this it is reasonable to neglect these variations since they are less significant than the exponential dependence of intrinsic carrier concentration on temperature.

From Equation 4.24,

$$n_i = N_c N_v \exp\left(\frac{-E_g}{2kT}\right) = N_c N_v \exp\left(\frac{-E_g}{2kT}\right)$$

In Example 1.2 we calculated the intrinsic carrier concentration in silicon at $T = 300$ K or $27°C$ as follows:

$$n_i = p_i = \sqrt{N_c N_v} \exp\left(\frac{-E_g}{2kT}\right)$$

$$= \sqrt{2.84 \times 10^{19} \times 1.06 \times 10^{19}} \exp\left(\frac{-1.12 \text{ eV}}{2 \times 0.026 \text{ eV}}\right) = 7.68 \times 10^9 \text{cm}^{-3}$$

If we repeat this at $-50°C$ (223 K) we obtain

$$n_i = \sqrt{N_c N_v} \exp\left(\frac{-E_g}{2kT}\right)$$

$$= \sqrt{2.84 \times 10^{19} \times 1.06 \times 10^{19}} \exp\left(\frac{-1.12 \text{ eV}}{2 \times 0.026 \text{ eV}} \times \frac{300}{223}\right)$$

$$= 4.52 \times 10^6 \text{ cm}^{-3}$$

and at $100°C$ (373 K) we obtain

$$n_i = \sqrt{N_c N_v} \exp\left(\frac{-E_g}{2kT}\right)$$

$$= \sqrt{2.84 \times 10^{19} \times 1.06 \times 10^{19}} \exp\left(\frac{-1.12 \text{ eV}}{2 \times 0.026 \text{ eV}} \times \frac{300}{373}\right) = 5.2 \times 10^{11} \text{ cm}^{-1}$$

Now from

$$I_0 = qA\left(\frac{D_n}{L_n}\frac{n_i^2}{N_A} + \frac{D_p}{L_p}\frac{n_i^2}{N_D}\right)$$

we obtain

$$\frac{I_{0(223 \text{ K})}}{I_{0(300 \text{ K})}} = \left(\frac{n_{i(223 \text{ K})}}{n_{i(300 \text{ K})}}\right)^2 = \left(\frac{4.52 \times 10^6 \text{ cm}^{-3}}{7.68 \times 10^9 \text{ cm}^{-3}}\right)^2 = 3.46 \times 10^{-7}$$

and

$$\frac{I_{0(373 \text{ K})}}{I_{0(300 \text{ K})}} = \left(\frac{n_{i(373 \text{ K})}}{n_{i(300 \text{ K})}}\right)^2 = \left(\frac{5.2 \times 10^{11} \text{ cm}^{-3}}{7.68 \times 10^9 \text{ cm}^{-3}}\right)^2 = 4.58 \times 10^3$$

(b) Now assuming the same illumination level as used in Example 4.1,

$$V_{OC(223\,K)} = \frac{kT}{q} \ln\left(\frac{I_L}{I_0} + 1\right)$$

$$= 0.026\left(\frac{223}{300}\right) V \ln\left(\frac{10.9\,A}{1 \times 10^{-9}A} \times \frac{1}{3.46 \times 10^{-7}} + 1\right) = 0.73\,V$$

and

$$V_{OC(373\,K)} = \frac{kT}{q} \ln\left(\frac{I_L}{I_0} + 1\right)$$

$$= 0.026\left(\frac{373}{300}\right) V \ln\left(\frac{10.9\,A}{1 \times 10^{-9}\,A} \times \frac{1}{4.58 \times 10^3} + 1\right) = 0.48\,V$$

To obtain efficiency we require the fill factor FF. Assuming that FF and short-circuit current I_{SC} remain unchanged from Example 4.1, the maximum room temperature output power at 300 K is

$$P(300\,K) = I_{MP}V_{MP} = FF \times I_{SC}V_{OC} = 0.75 \times 10.9\,A \times 0.601\,V = 4.91\,W$$

At $-50°C$ (223 K),

$$P(223\,K) = I_{MP}V_{MP} = FF \times I_{SC}V_{OC} = 0.75 \times 10.9\,A \times 0.73\,V = 6.0\,W$$

and at 373 K

$$P(373\,K) = I_{MP}V_{MP} = FF \times I_{SC}V_{OC} = 0.75 \times 10.9\,A \times 0.48\,V = 3.9\,W$$

This clearly demonstrates the role of temperature on solar cell efficiency. The room temperature efficiency increased by roughly 20% at $-50°C$ and decreased by about 20% at 100°C.

Actual measured temperature coefficients of commercial silicon solar cells are in the range of 0.4% K^{-1}, which is equivalent to roughly a 30% increase at $-50°C$ and a 30% decrease at 100°C. These temperature coefficients are only valid near room temperature.

Note that solar cells made using semiconductors with larger bandgaps have less temperature dependence because n_i is less temperature dependent. See Chapter 1.

4.8   Silicon Solar Cell Technology: Wafer Preparation

The most widely manufactured solar cells are based on the use of silicon. There are three main types of silicon solar cells materials. They are *single crystalline* silicon, *multicrystalline* silicon and *amorphous* silicon. Each material has a strong niche in the solar cell market, but the performance levels and other attributes differ dramatically. The manufacture and design aspects of silicon-based solar cell types will be reviewed in this chapter, and solar cells made using other materials will be covered in Sections 4.11–4.13.

Single crystal silicon cells have efficiency values that can exceed 20% in production. The starting materials must be highly purified. To achieve the highest purity silicon, naturally occurring quartz (SiO_2) is reduced by the reaction

$$SiO_2 + 2C \rightarrow Si + 2CO$$
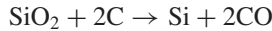
The reaction occurs in a furnace in which carbon in the form of coke is introduced along with the SiO_2 at temperatures well above $1414°C$, which is the melting point of silicon. In addition, to react and remove aluminium and calcium in the silicon, oxygen is blown into the furnace. The resulting *metallurgical grade* silicon, or *MG silicon*, is 98–99% pure and typically contains atomic percentages of metal impurities in the range of 0.3% Fe, 0.3% Al, 0.02% Ti and under 0.01% of each of B, Cr, Mn, Ni, P and V. There is also some residual oxygen incorporated.

To achieve semiconductor grade silicon, further purification is required. The standard process is the *Siemens process*. Fine MG silicon powder is reacted with HCl to produce gaseous trichlorosilane through the reaction

$$Si + 3HCl \rightarrow SiHCl_3 + H_2$$
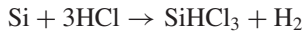
The trichlorosilane (melting point $-126.2°C$, boiling point $31.8°C$) is condensed to a liquid and then distilled several times to upgrade its purity. Upon reacting it with hydrogen at elevated temperature, the reduction reaction

$$SiHCl_3 + H_2 \rightarrow Si + 3HCl$$
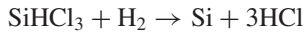
takes place at the hot surface of a high-purity silicon rod. The rod is continuously coated with the resulting semiconductor grade silicon in the form of a fine-grained polycrystalline deposit. Eventually thicknesses of silicon in the range of 10 cm may be deposited onto the rod. The purity available is in the parts per billion range, and the semiconductor industry relies on this process for its supply of silicon.

This process is high in energy consumption, involves environmentally hazardous chloride chemistry and adds significantly to the ultimate cost of silicon solar cells. In spite of this, it remains the dominant process used for the final purification of standard solar grade silicon in the industry.

There are lower cost processes that have been developed to partially purify MG silicon. For example, MG silicon may be mixed with aluminium and a molten solution of silicon and aluminium may be cooled to precipitate purified crystalline platelets of silicon having purity in the parts per million range. The aluminium phase may be removed from the platelets by melting it and pouring it off, followed by an acid washing step in which remaining aluminium is dissolved away. Further purification can then be carried out as required.

The purity requirement of solar grade silicon is lower than that for semiconductor grade silicon, and the purification process may be optimized to provide adequate solar cell performance. This depends on the desired specifications of the solar cells, which range from 14% to 24% efficiency. The precise optimization of the purification process is a highly competitive aspect of solar cell manufacturing and is typically not disclosed by manufacturers. The resulting silicon is called *solar-grade* material. Regular measurement of minority carrier lifetime during production is an important and very useful probe of the impurity level achieved since this lifetime is highly sensitive to impurity concentration. See Section 2.8.

The highest efficiency silicon solar cells are single crystal. Solar grade silicon is melted and a suitable concentration of boron is intentionally added to produce the p-type silicon
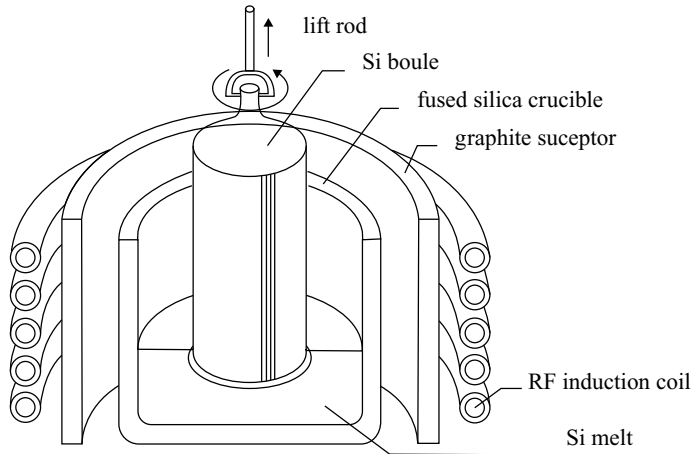
Figure 4.17   *Czochralski growth system. http://electricly.com/tag/semiconductor/. Reprinted from http://electricly.com/tag/semiconductor/. Copyright © 2010 Free Electronics and Engineering Tutorial*

required for the thick p-type layer of the solar cell of Figure 4.7. A single crystal silicon seed is lowered to the liquid surface and slowly pulled from the melt, resulting in the growth of a single crystal *boule* or circular rod of single crystal silicon having a diameter in the 15 to 20 cm range and a length of 1 to 2 metres. This is known as the *Czochralski* growth process (see Figure 4.17).

The boule is then sliced into silicon wafers that are approximately 200 μm (0.2 mm) thick using a *wire saw* process. An array of closely spaced, tensioned metal wires is continuously pulled against the side of a silicon boule in the presence of an abrasive slurry. The rubbing action of the wire against the boule wears away the silicon and a series of cuts is made. The kerf of these cuts is primarily determined by the diameter of the wire and may be in the range of 200 μm. Hundreds of silicon wafers may be cut at one time as the array of wires slowly passes through the boule. The advantages of this process include the low stress caused by any one wire, which prevents damage to the silicon, and the ability to prepare a large number of wafers per cycle. The steady replacement of the abrasive slurry is provided by a pumping system, which ensures that the cutting speed is maintained. A disadvantage of this cutting process is the loss of approximately 50% of the available silicon. In spite of this loss the wire saw approach is the predominant silicon wafering method and it is improving steadily. Over the past decade the ability to cut thinner wafers has progressed through improvements to all aspects of this process. Wafers have reduced in thickness from 300 μm to under 200 μm, and there is an opportunity to reduce this further to approach the limiting 100 μm thickness required for absorption of sunlight in silicon. Slurries containing diamond abrasives can further improve cutting speed and quality. It is also interesting to note that whereas the semiconductor industry currently uses 30 cm diameter boules of silicon and wafer thicknesses of 300–600 μm, the solar cell industry has maintained its use of smaller diameter boules since they assist in the reduction of wafer thickness and also prevent excessive electric currents being generated in a single silicon wafer.

The market share of multicrystalline silicon in solar cells has grown to rival the use of single crystal silicon. This is a result of cost reduction associated with the elimination
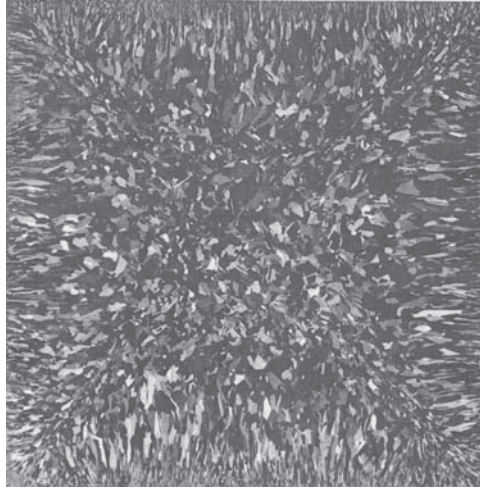
Figure 4.18 *A 10 × 10 cm multicrystalline silicon wafer. Note the large single-crystal grains that can be up to approximately 1 cm in dimension. Reprinted from Green, M. A., Solar Cells p. 128, 130, 189 ISBN 0-85823-580-3. Copyright (1981) with permission from M. A. Green*

of the Czochralski growth process as well as a steady improvement in the efficiency of multicrystalline silicon solar cells, which are typically in the range of 14–18% efficient versus 16–24% for single crystal silicon. Rather than pulling a single crystal from the melt, a casting process is used in which molten solar-grade silicon is poured into a mould. Solidification takes place in a very slow and controlled manner to optimize the growth of very large-grained silicon. Grain sizes in the range of 1 cm may be achieved. Square cross-section ingots of multicrystalline silicon formed in this manner may be sliced using wire saws to produce square silicon wafers. The square wafers may be assembled into solar cell modules more efficiently than the wafers from the single crystal process. After cutting, the wafers may be polished and are then ready for cell fabrication steps. Multicrystalline silicon is shown in Figure 4.18.

The grain boundaries of multicrystalline silicon must be treated to make them behave like low recombination velocity surfaces. In the p-type material it is therefore appropriate to introduce p^+ doping into the grain boundaries. This forms the equivalent of a back surface field at each grain boundary. Elements such as boron or aluminium can be used provided they are preferentially introduced into the grain boundaries of the multicrystalline silicon by either grain boundary diffusion or grain boundary segregation.

4.9 Silicon Solar Cell Technology: Solar Cell Finishing

An important and widely used first step achieves texturing of the silicon surface. If the silicon wafers are cut with (100) orientation, a selective etch may be used to achieve a surface covered with square-based pyramids of silicon, as shown in Figure 4.19. These pyramids are typically 10 μm high and improve the light absorption of the silicon surface. Light reaching the silicon surface may enter the solar cell through the pyramid side facets or it may reflect off the side facet of one pyramid and enter the solar cell through another pyramid as shown.
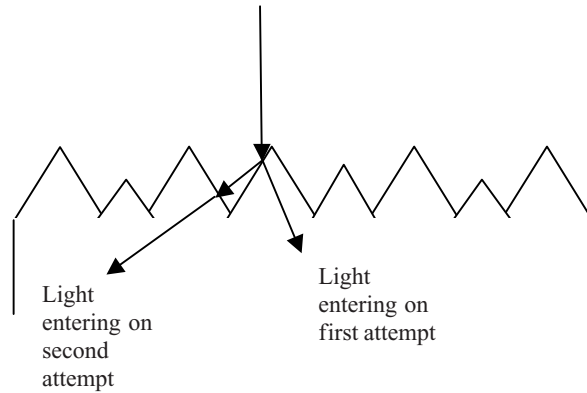
Figure 4.19   *The possible paths of light beams reaching the solar cell surface are shown. At least two attempts to enter the silicon are achieved for each light beam. The angle of the sides of the pyramids may be calculated from the known crystal planes of silicon*

Silicon solar cells are bonded behind low-iron-content, high-transmission solar grade glass using transparent polymer bonding material that is capable of filling in voids between the silicon and the glass. Iron impurities found in standard float glass absorb sunlight and solar grade glass therefore has lower iron content to improve solar panel efficiency. The reflectivity of sunlight off an untreated silicon surface embedded in a polymer behind a glass plate is approximately 20%. If, on average, every reflected light beam is directed back to the silicon surface once again, as shown in Figure 4.19, then the reflectivity is reduced to $20\% \times 20\% = 4\%$, which is achieved in practice. Even lower reflectivity may be obtained using an antireflection coating; however, this would be applied after other fabrication steps are complete. A micrograph of pyramids formed in an etched silicon surface is shown in Figure 4.20.

Once the desired texturing is complete the front surface of the wafer must be doped n-type. This requires a phosphorus diffusion, which may be achieved using commercially available printable paste containing phosphorus, which is applied using a *screen printing* process to the front surface of the wafer. A paste having the consistency of syrup is deposited on the surface of a fine mesh or screen. The silicon wafer is placed just below but not touching the screen, and a flexible blade is passed over the screen forcing contact between the wafer and the screen and causing a well-controlled layer of paste to pass through the screen and onto the wafer. After a low-temperature bake the volatile components of the paste are released leaving only the desired phosphorus on the silicon. A high-temperature diffusion step is then used to diffuse the phosphorus to the desired depth in the silicon, which is typically set to form the p-n junction about 1 μm below the surface. The result is an n^+-layer with $N_d > 10^{18}$ cm^{-3} near the front surface.

The back contact is achieved using an aluminium dopant that forms a p^+ doped layer. This layer serves two functions. It creates the back surface field for a low recombination velocity back surface, and it also allows for a tunnelling-type ohmic contact to the back aluminium metal contact. Commercially available aluminium pastes are available that contain the aluminium in a paste suspension. A low-temperature bake volatilizes the unwanted components and subsequently the aluminium is diffused into the silicon in a high-temperature anneal.
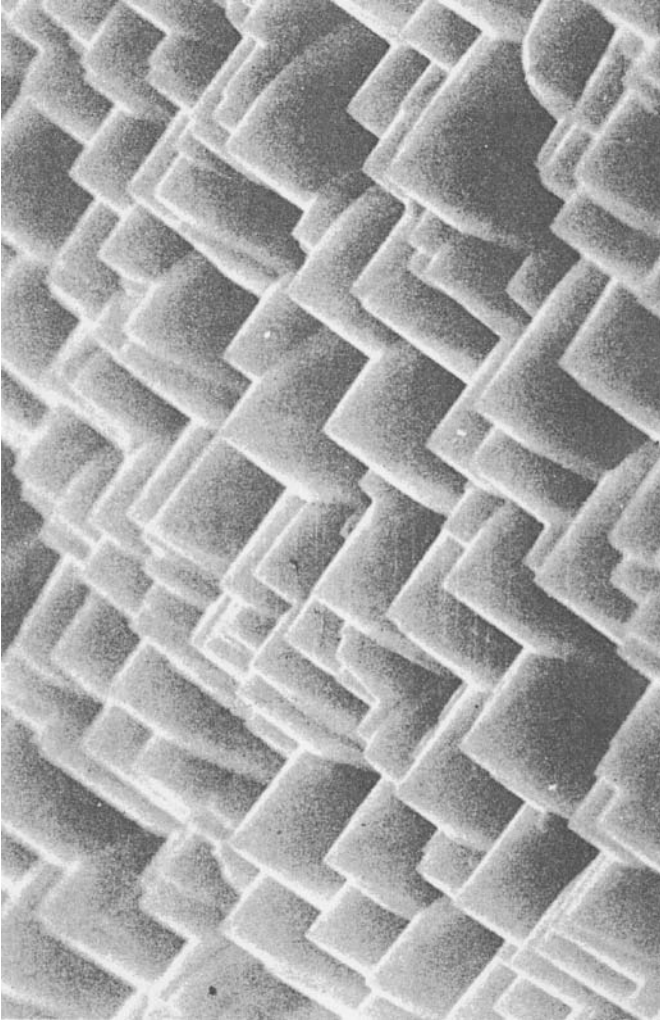
Figure 4.20 *Micrograph of pyramids selectively etched in the surface of a (100) silicon wafer. Sides of the pyramids are {111} planes, which form automatically in, for example, a dilute NaOH solution. Note the scale showing that the pyramids are approximately 10 μm in height although there is a range of pyramid heights. The pyramids are randomly placed on the surface; however, they are all orientated in the same direction due to the use of single-crystal silicon. Reprinted from Green, M. A., Solar Cells p. 128, 130, 189 ISBN 0-85823-580-3. Copyright (1981) with permission from M. A. Green*

Silver paste containing a small percentage of aluminium is also available, which forms a silver back metallization. This permits soldering to the back of the solar cell during module manufacture. The screen printing method allows the silver paste to be deposited selectively onto small areas of the back of the wafer, which is done to enable soldering while minimizing the amount of silver required. If the screen is only porous in the areas in which
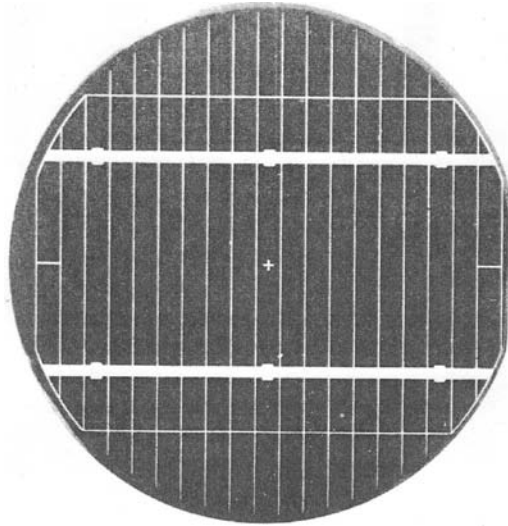
Figure 4.21   *Pattern of conductors used for the front contacts of a silicon solar cell showing bus bars as well as the narrow conducting fingers. The electrode was screen printed. Reprinted from Green, M. A., Solar Cells p. 128, 130, 189 ISBN 0-85823-580-3. Copyright (1981) with permission from M. A. Green*

a deposit is desired, then when the rubber blade is passed over the screen during printing, paste is only applied to the silicon in these areas. Screens may be prepared with a spatially patterned polymer masking layer that blocks the pores of the screen in selected areas. A suitably masked screen may be used to print a given pattern onto thousands of wafers.

The front contact of the solar cell may now be printed. This requires a set of narrow conductors that are typically connected by bus bars, as shown in Figure 4.21. Since light loss increases with front contact area, conductor width must be minimized and is routinely in the range of 200 μm. Spacing between the conductors is determined by the sheet resistance of the n^+ layer. The sheet resistance depends upon the phosphorus doping level, which is optimized to avoid excessive doping levels that exceed the solubility limit of the dopant in the silicon. The thickness of the n^+ layer must also be minimized to reduce light absorption in the n^+ silicon. Since the front surface recombination velocity of the solar cell is typically high, it is better to absorb the bulk of the light in the depletion layer and in the p-type silicon layer.

As a result there is an upper practical limit to the conductivity achievable in the n^+ layer. This is defined by the *sheet resistance R_s* of the layer, expressed in units of ohms, where

$$R_s = \frac{\rho}{T}$$

ρ and T being layer resistivity and layer thickness respectively. R_s is the resistance of a square unit of the layer measured across conductive contact strips running along two opposing edges of the square unit. The value of R_s determines the spacing between the narrow conductors. A typical spacing is approximately 3 mm, as shown in Figure 4.21. An effective method of forming these contacts is through the use of silver-based paste containing n-type dopant phosphorus to ensure a good ohmic contact. The paste is screen printed through a suitable masked screen. The narrow conductors and bus bars may then be

printed in one step. As with the back contact metallization, low- and then high-temperature annealing steps are performed.

An optional additional antireflection layer applied to silicon solar cells should have a thickness of $\frac{\lambda}{4}$ and a refractive index of $\sqrt{n_{Si}\, n_{polymer}}$ where n_{Si} and $n_{polymer}$ are the indices of refraction of the silicon and the polymer bonding material. A suitable material is TiO_x, which is usually deposited by sputtering.

The solar cell is now ready for soldering to conductors. Bonding and packaging materials for cell mounting onto a front glass substrate have been well developed for this process. A completed solar module containing multiple wafers and capable of withstanding the outdoor environment for over 25 years is the result.

4.10   Silicon Solar Cell Technology: Advanced Production Methods

Since by far the majority of solar cells are made from crystalline silicon, it is not surprising that a variety of production techniques have been developed and implemented in order to lower the cost of production.

Single crystal silicon continues to lead in silicon solar cell efficiency. This is partly due to a purification process that is inherent in the Czochralski growth process: impurities in the molten feedstock segregate into the remaining liquid silicon since the liquid has a higher solubility for impurities than the crystalline material. The crystallized silicon is therefore higher in purity than the feedstock. The most important impurity is generally iron, and impurity levels below 10^{12} atoms cm^{-3} of iron are required to achieve carrier diffusion lengths well above $150\,\mu m$. The speed of Czochralski growth may be increased beyond the speed used in the microelectronics industry to values in the range of 1 mm per minute. The resulting carrier diffusion length is carefully monitored and process speed is optimized to minimize cost. The simplicity of the growth method allows for highly repeatable and dependable growth of high-quality silicon. In multicrystalline silicon typical ingot sizes have reached over 250 kg. Wafer thicknesses of approximately $200\,\mu m$ are standard. In polycrystalline solar cells the casting of silicon does not permit this purification process and the purity of the feedstock is more carefully controlled.

Although the boule and ingot-based methods of silicon wafer production are well developed and are the dominant methods of production of silicon solar cells, they both require sawing with the associated material loss. An interesting alternative to this is *silicon ribbon technology*, which results in the direct solidification of molten silicon into thin silicon sheets or ribbons. The best known approach is called the *string ribbon* growth method. Two parallel strings separated by several centimetres are pulled vertically through a crucible of molten silicon. As the liquid wets the strings, a thin web of liquid is maintained between the strings through surface tension, which solidifies into a polycrystalline silicon sheet upon cooling, as shown in Figure 4.22.

Challenges associated with silicon ribbon growth include:

(a) removal of the strings by cutting off the edges of the ribbon;
(b) the degree of flatness achievable in production;
(c) the ultimate solar cell performance;
(d) impurity incorporation due to the molten silicon in contact with its container;
(e) the energy used in maintaining molten silicon temperatures during pulling.
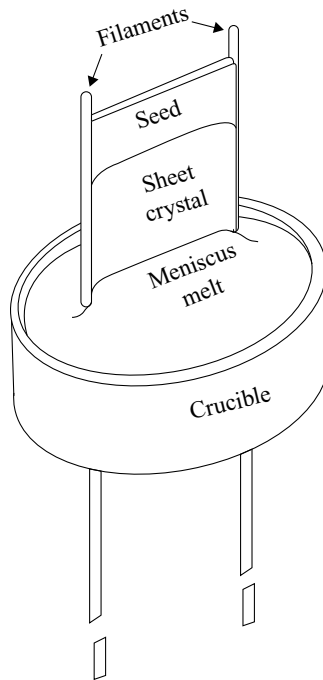
Figure 4.22   *String ribbon growth of a silicon ribbon. Surface tension in liquid silicon forms a silicon sheet upon cooling. Thickness of the ribbon is controlled by pull rate and rate of cooling of the silicon web. Reprinted from Luque, A., and Hegedus, S., Handbook of Photovoltaic Science and Engineering, ISBN: 978-0-471-49196-5. Copyright (2003) with permission from John Wiley & Sons*

Cell efficiency is in the range of 14%, which is lower than wafered cells due to the incorporation of impurities and the multicrystalline nature of the ribbon. Nevertheless the string ribbon method is a proven production process.

4.11   Thin Film Solar Cells: Amorphous Silicon

There are strong driving forces to lower production costs of solar cells. There remains a significant cost penalty to the use of solar cells compared to traditional fossil fuels and nuclear power. The most important approach to lowering cost is thin-film PV. Rather than bulk or ribbon growth of the active material, thin films of semiconductor layers are deposited on a substrate that supports the thin films. The substrate is low in cost and becomes an integral part of the finished product.

In the early 1970s the semiconductor properties of a new form of silicon were discovered. Known as *amorphous silicon*, often written a-Si, this is a disordered form of silicon, generally prepared in thin-film form. The normal long-range bonding in silicon becomes disordered or random; however, local bonding remains largely tetrahedral as in single-crystal material. This a-Si material may be grown in thin-film form onto a supporting substrate.
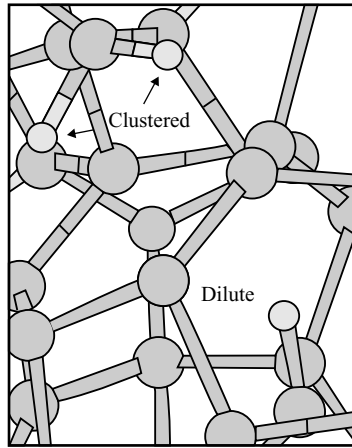
Figure 4.23 *Atomic structure of a-Si:H in which H atoms terminate dangling Si bonds, which are generally isolated but may also be clustered (two dangling bonds in one silicon atom). Reprinted from Luque, A., and Hegedus, S., Handbook of Photovoltaic Science and Engineering, ISBN: 978-0-471-49196-5. Copyright (2003) with permission from John Wiley & Sons*

Several percent of the tetrahedral silicon bonds are incomplete or *dangling* due to the lack of long-range order. If a suitable amount of hydrogen is incorporated with the a-Si thin film, hydrogen atoms effectively complete or *terminate* these silicon bonds. The optimum incorporation of hydrogen turns out to be achieved using a growth technique known as plasma enhanced chemical vapour deposition (PECVD), in which a gas mixture of SiH_4 (silane) and H_2 is subjected to an RF (radio frequency) discharge, which breaks the molecules apart to form radicals. These radicals then deposit onto a substrate placed in the discharge resulting in the growth of the desired thin film on the substrate. The resulting hydrogen-terminated film is known as a-Si:H material. The atomic structure of a-Si:H is shown in Figure 4.23.

It is of significance that the absorption of sunlight in a-Si:H is very different from the absorption in crystalline silicon. Figure 4.24 compares the absorption spectrum of single-crystal silicon with that in a-Si:H material. The effective bandgap of a-Si-H is higher than that of crystalline Si, which is evident from the solid curve that rises rapidly in an energy range about 0.5 eV higher compared to the dashed curve in Figure 4.24. Above 1.9 eV, however, the absorption coefficient in a-Si:H becomes higher by over one order of magnitude compared to crystalline silicon, which means that a thin film of a-Si:H in the thickness range of 1 μm is enough to absorb considerable sunlight. This may be understood from a band model for a-Si:H in which the indirect gap properties of silicon are relaxed by the lack of long-range order, which makes a:Si behave somewhat more like a direct gap semiconductor.

Interesting features of the energy band diagram for amorphous silicon are shown in Figure 4.25. Due to long-range disorder the values of conduction band edge E_c and valence band edge E_v are not clearly defined as in crystalline silicon. This occurs because small variations in bonding energy exist for different bond angles and configurations associated with the structure in Figure 4.24. Density of states functions exist both for
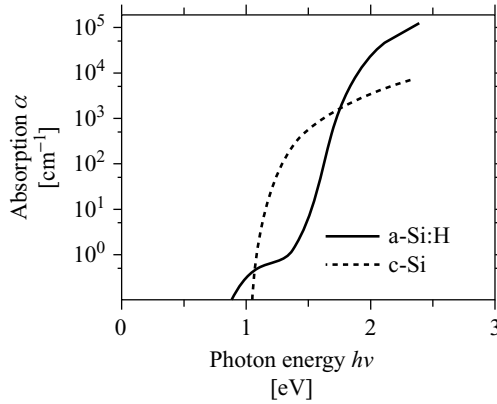
Figure 4.24 *Absorption edge of amorphous silicon compared to crystalline silicon. There is about a half electronvolt energy difference in absorption edge between the two materials. Reprinted from Luque, A., and Hegedus, S., Handbook of Photovoltaic Science and Engineering, ISBN: 978-0-471-49196-5. Copyright (2003) with permission from John Wiley & Sons*

electrons below E_c and for holes above E_v that exponentially decrease in density into the energy gap. These are called *bandtails*. The bandgap is therefore not precisely defined, but may be approximated as in Figure 4.25 to be about 1.75 eV, although it does depend upon the preparation conditions of the material. There is an important distinction between states in bandtails within the bandgap and states outside the bandgap: the former are
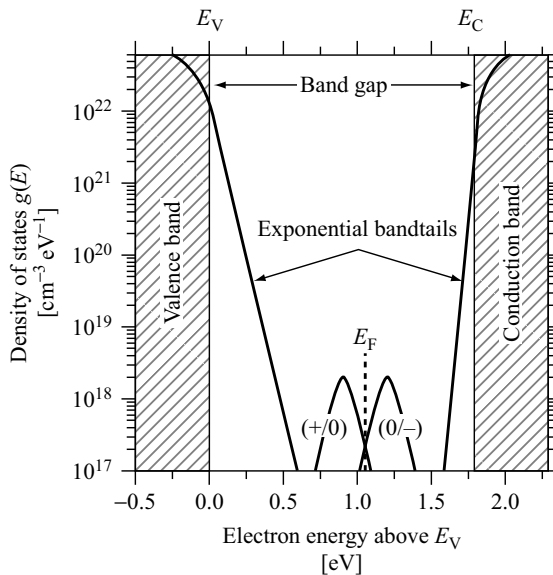
Figure 4.25 *Density of electron states in a-Si:H. Note the bandtails as well as the mid-gap states due to defects. Reprinted from Luque, A., and Hegedus, S., Handbook of Photovoltaic Science and Engineering, ISBN: 978-0-471-49196-5. Copyright (2003) with permission from John Wiley & Sons*

localized whereas the latter are delocalized in ideal crystalline semiconductor material, all the band states are delocalized. Bandtails arise from localized disorder, which results in localized states. Conduction in a-Si:H is essentially due to the delocalized states outside the bandgap.

There are also states closer to mid-gap. These are states that arise due to defects such as dangling silicon bonds. Dangling bonds were discussed in the context of surfaces and interfaces in Chapter 1, but they can also exist in a-Si:H throughout the semiconductor. The density of these defects is strongly related to the degree of hydrogen termination of dangling bonds. As-prepared material may have a low defect density below $10^{16}\,\mathrm{cm}^{-3}$. Unfortunately extended exposure to sunlight is well known to increase the defect density to about $10^{17}\,\mathrm{cm}^{-3}$, which degrades the performance of solar cells over time until the new, higher defect density is stabilized. This process is known as the *Staebler–Wronski effect*, and is explained by the rearrangement of hydrogen atoms within the a-Si:H.

Doping in a-Si:H may be achieved by the incorporation of impurities such as phosphorus and boron as in crystalline silicon. In n-type a:Si:H the measured free electron concentration is much smaller than the P doping level. This may be explained because a significant fraction of doped P atoms will occupy sites that are only bonded to three nearest silicon neighbours and the extra two P valence electrons remain in pairs tightly attached to the P atom. This does not lead to an additional localized shallow electron state and therefore does not result in n-type doping. Occasionally, however, the P atom occupies a site that is bonded to four nearest silicon neighbours and n-type doping is achieved.

Because there is a relatively high density of electron pairs bound to inactive P atoms in n type a-Si:H, this material does not allow for effective hole transport because holes are trapped by these electron pairs. This puts constraints on the design and structure of an effective a-Si:H solar cell.

The best structure for an amorphous silicon solar cell is suitably named the *p-i-n* structure. This is a p-n diode with a thick intrinsic layer sandwiched between thin p- and n-layers. Typical devices have $\cong 20\,\mathrm{nm}$ thick p- and n-layers and a $\cong 500\,\mathrm{nm}$ thick intrinsic layer. The reason for the insertion of the intrinsic layer relates to the problem of achieving both adequate hole and electron mobility in doped a-Si:H material. The goal is to ensure that virtually all the photon absorption occurs in the intrinsic layer. This is in contrast to the crystalline solar cell, in which most absorption occurs in the p-type material. Mobilities of approximately $1\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ are typical in the intrinsic layer. This mobility value is small compared to crystalline silicon carrier mobility of order of magnitude $1000\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$; however, the typical thickness of amorphous solar cells of $0.5\,\mu\mathrm{m}$ is much thinner than a typical crystalline cell thickness of $200\,\mu\mathrm{m}$. This allows relatively small mobilities to be adequate.

The intrinsic layer in the thin film solar cell effectively establishes the width of the depletion region in the cell. The built-in electric field is established across this intrinsic layer, and a key requirement is that carriers optically generated in this layer should reach the appropriate n- and p-layers before recombining. The magnitude of this built-in field can be estimated as follows: a 500 nm layer in a built-in potential of 0.5 V results in an electric field of $10^4\,\mathrm{V}\,\mathrm{cm}^{-1}$. Using $v = \mu\varepsilon$ and taking $\mu = 1\,\mathrm{cm}^2\,\mathrm{V}^{-1}\,\mathrm{s}^{-1}$ and $\varepsilon = 10^4\,\mathrm{V}\,\mathrm{cm}^{-1}$ we obtain $v = 10^4\,\mathrm{cm}\,\mathrm{s}^{-1}$. If the intrinsic layer is $L = 5 \times 10^{-5}\,\mathrm{cm}$ in thickness the carrier transit time to cross over the intrinsic layer becomes $L/v = 5$ nanoseconds, and is is clear that carrier lifetimes in amorphous silicon can be much shorter than carrier lifetimes in the microsecond range required for crystalline silicon solar cells.

The open circuit voltage of amorphous silicon cells is higher than in crystalline cells by about half a volt. This is explained by the increase in the absorption edge by a corresponding 0.5 eV in Figure 4.23. Typical measured open circuit voltages are close to 1 V.

There are two basic a-Si:H device designs in which illumination is either incident through the substrate onto the lower surface of the cell or directly incident onto the upper surface of the cell.

In the transparent substrate design, a glass substrate is used and is coated first with a transparent electrode composed of a transparent conductive oxide (TCO) material such as tin-doped In_2O_3 (ITO) or aluminum-doped ZnO followed by the p-type layer, which acts as a window, the intrinsic layer, the n-type layer and finally a rear electrode. It is usual for this rear electrode to be reflective since it allows light that was not absorbed to reflect back and once again have an opportunity to be absorbed. This design lends itself to low cost rigid solar panels. In the direct illumination design, the substrate may be non-transparent. A practical material is a thin stainless-steel sheet, which serves as the rear electrode and yields a flexible solar cell. The deposition sequence includes an n-type layer, the intrinsic layer and then the p-type layer followed by a transparent front electrode. This type of design features bendability and low weight. See Figure 4.26.

An opportunity exists to further increase the percentage of light absorbed in the structures of Figure 4.26. This is accomplished using back reflectors that are not planar, but light scattering instead. Light that has traversed the active layers once has a high chance to reflect off the back reflector and back into the active layers at high angles relative to the thin-film normal axis, which means that the optical path length through the active intrinsic layer is higher. In addition, light that has made a round trip and reaches the front electrode
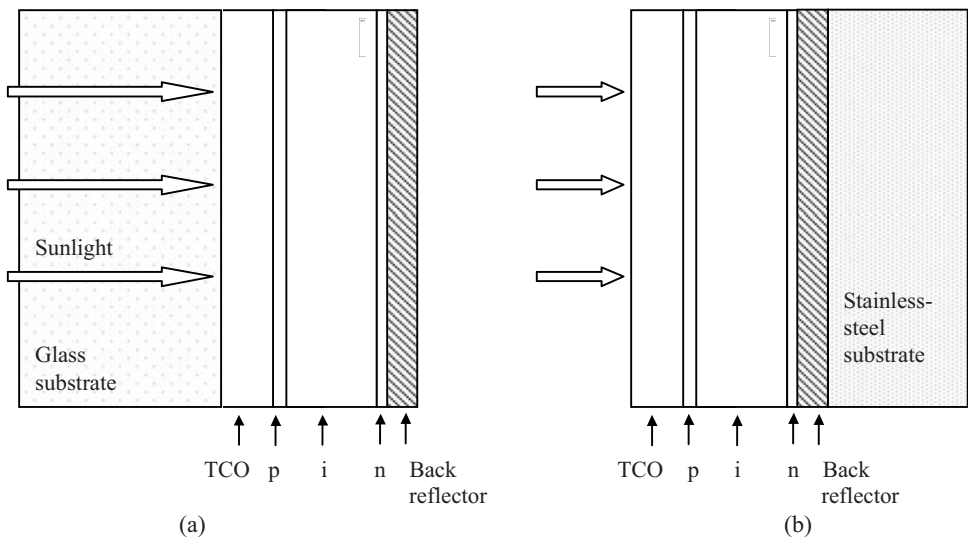
Figure 4.26   *Structures of a-Si:H solar cells (not to scale). (a) Glass substrate structure with illumination through the substrate. (b) Stainless-steel substrate structure with direct illumination. TCO, transparent conductor layer*

may exceed the critical angle and again reflect by total internal reflection and remain trapped inside the solar cell. Texturing of reflectors and TCO (transparent conducting) layers is used to maximize conversion efficiency in commercial devices.

The commercially achievable conversion efficiency of single-junction a-Si:H solar cells is approximately 7% for un-aged devices and approximately 5% after 1000 hours of exposure to sunshine, which allows the Staebler–Wronski degradation to stabilize. This is three to four times lower than crystalline silicon solar cell efficiency values. The application of single-junction amorphous silicon solar cells is therefore limited to applications where low power and low cost are needed.

A great deal of work has been done to increase the conversion efficiency of a-Si:H solar cells, and the most effective approach is to form a *multiple junction* device. Two or more thin-film p-i-n junctions are stacked and effectively connected in series. Light that is not absorbed in the first junction passes to the second junction, and so on. The total voltage becomes the sum of the open circuit voltages from each p-i-n junction.

The key to success is to change the energy gap of each junction such that the first junction has the highest energy gap, which absorbs high-energy photons but transmits lower energy photons to the second junction, which has a lower energy gap, and so on. In this scheme the open circuit voltage of the first p-i-n junction is highest, and takes better advantage of the higher energy photons, whereas in a single-junction solar cell with a smaller bandgap these high-energy photons generate electron-hole pairs that lose their excess energy in the form of heat before they get collected. This concept is known as *spectrum splitting*.

Tandem cells have two p-i-n junctions, and *triple-junction* cells have three p-i-n junctions. The maximum amount of energy available increases with the number of cells in the stack since carrier energy loss through thermalization decreases as the number of junctions increases. The optimum energy gaps required depend on the solar spectrum. For tandem cells an optimized stack included a first gap of $\cong 1.8\,\text{eV}$ and a second gap of $\cong 1.2\,\text{eV}$.

There are other benefits that arise through the use of multiple junction cells. Total current for a given electrical output power decreases due to the higher output voltage. This decreases resistive losses. Also the thickness of each junction is less than for a single-junction device. This lowers carrier recombination losses and effectively increases the fill factor of the overall device.

In order to achieve the required bandgap adjustments, Si must be combined with other elements during the formation of the amorphous layers. The best understood alloy is a-$Si_{1-x}Ge_x$:H. By changing the value of x the bandgap may be adjusted from 1.1 eV ($x = 1$) to 1.7 eV ($x = 0$). This would appear to be an almost ideal way to prepare a tandem device; however, there are challenges. Germanium is much more expensive than silicon, and more importantly the effective defect density in a p-n junction increases as x increases, which lowers the fill factor of these p-n junctions. The defects trap carriers (holes in particular) in the intrinsic layer preventing them from being collected. In practice the minimum achievable bandgap is 1.4 eV.

One interesting way to reduce this problem is to vary the germanium concentration as a function of depth within the intrinsic layer of a given p-n junction. If the germanium content is raised towards the p-side of the junction then holes photogenerated in the i-layer will have a shorter distance to travel to the p-side, effectively decreasing their chance of being trapped.
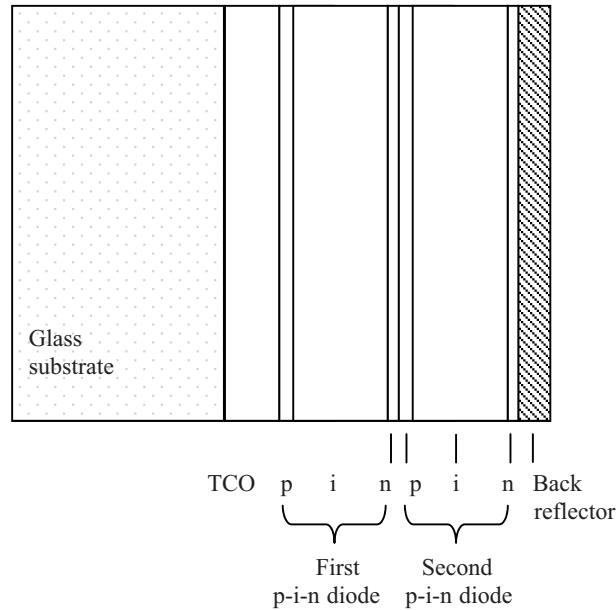
Figure 4.27 *Tandem solar cell structure on glass substrate. The n-p junction formed at the interface between the first and second p-i-n diodes must be an effective tunnelling junction to allow carriers to flow to the next diode. The bandgap of the first p-i-n diode is higher than the bandgap of the second p-i-n diode. Similar tandem and triple junction structures may be formed on stainless-steel substrates. TCO is a transparent conductive oxide layer*

The most efficient practical triple junction amorphous cells are a-Si(1.8 eV)/a-SiGe($\cong$ 1.6 eV)/a-SiGe($\cong$ 1.4 eV), and after stabilization over 10% efficiency is commonly achieved. To reach these efficiency values the operating points shown in Figure 4.9 must be matched in all three junctions. Specifically I_{MP} for each junction must be the same under normal sunlight conditions. This requires careful optimization of the thicknesses of each p-i-n junction. Back reflectors are incorporated, which influence the optimum layer thicknesses of each junction.

A key requirement for the successful operation of multiple junction solar cells is adequate conduction between each junction in the stack. In Figure 4.27 the structure of a tandem cell is shown. The n-p junction formed between the two p-i-n junctions is *reverse* biased during normal operation; however, it does allow current to flow because it operates as described for a tunnel diode in Figure 2.21. Tunnelling current can flow, provided degenerate doping levels are applied to both the n- and p-layers, and these layers are therefore very heavily doped. Fortunately this does not compromise the operation of the p-i-n devices since the critical light-absorbing layer is the intrinsic layer.

There is also a development in silicon thin-film multiple junction solar cells in which one back layer of silicon is crystallized and is referred to as *polycrystalline* or *microcrystalline* rather than amorphous. These cells may be grown on glass substrates and crystallization is achieved by *rapid thermal annealing*. The polycrystalline silicon layer has an absorption

edge that is lower in energy allowing better utilization of long wavelengths. However, the challenges are the following:

(a) Owing to the smaller absorption coefficient of crystalline silicon the recycling of light is critical, and highly reflective electrodes and light-trapping schemes are essential.
(b) The polycrystalline silicon presents a large number of grain boundaries that act as recombination centres. The density of these grain boundaries can be decreased by increasing the silicon grain size, and grains of several microns are achievable. Increasing the grain size requires novel growth methods to avoid processing temperatures that exceed the glass softening temperature. Rapid thermal annealing may be performed using *laser recrystallization*, in which rapid heating of the silicon without melting the glass substrate is possible In addition a suitable grain boundary surface treatment must be performed to minimize grain boundary surface recombination velocity.

Silicon tandem solar cells based on two amorphous silicon layers as well as one micro-crystalline silicon layer now yield over 10% conversion efficiency in mass production and constitute an important development in solar cell evolution due to their inherently low cost of manufacture and absence of toxic elements. See Problem 4.11.

4.12   Telluride/Selenide/Sulphide Thin-Film Solar Cells

The best known large-area thin-film solar cells for power generation are not silicon based. Instead the group II-VI semiconductor CdTe has been exploited, and thin-film cells with over 11% conversion efficiency are in large-scale production. More recently strong interest in a selenide semiconductor material $Cu(InGa)Se_2$ (CIGS) is yielding thin-film solar cells with over 13% conversion efficiency.

CdTe is a direct-gap semiconductor with a bandgap of 1.5 eV. Since it is a direct-gap material it has a high optical absorption coefficient of 5×10^5 cm^{-1} for photons with energy greater than 1.5 eV. CIGS is also a direct-gap material. See Figure 1.16f.

The form of these semiconductors used in thin-film devices is polycrystalline. For CdTe cells a thin film of p-type CdTe of about 2 μm in thickness is grown on a substrate such as glass. The film is composed of a large number of columnar single crystal grains. The size of a grain is typically in the range of 0.2–1.0 μm in cross-section, and grain height is the same as the film thickness. In addition to the p-type CdTe layer there are other thinner layers in CdTe solar cells including an n-type CdS layer and a TCO layer. CdS has a wider bandgap than CdTe and therefore acts simultaneously as a window material that offers minimal light absorption and as the n-side of the p-n junction. The CdTe/CdS structure is an example of a heterojunction solar cell (see Figure 4.28).

CdTe homojunction solar cells can be formed, and CdTe may be grown either n-type or p-type. One method of achieving n-type CdTe is to introduce excess Cd, whereas for p-type CdTe excess Te is used. Dopants such as In and Au can also be used to achieve n-type and p-type doping respectively; however, homojunctions have not demonstrated the highest solar cell efficiencies.

A large number of heterojunction solar cells have been investigated. In some of these the CdTe was used for the n-type layer. In addition p-n junctions using p-type single-crystal
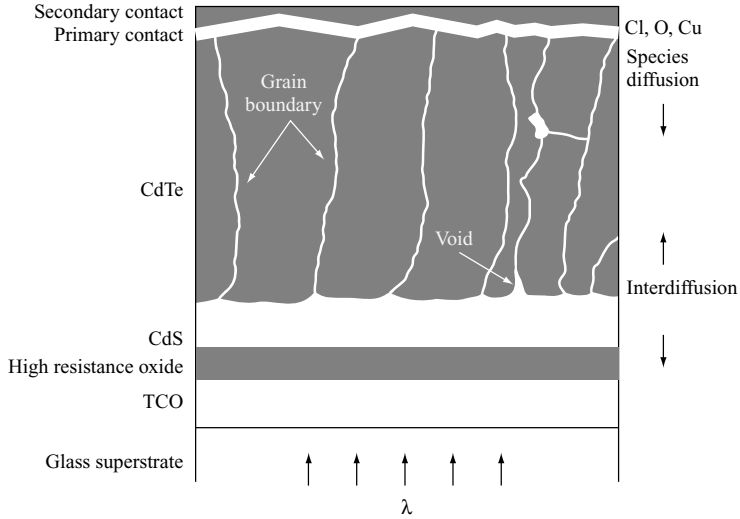
Figure 4.28   *Grain structure of CdTe solar cell also showing the CdS layer and device structure. TCO, transparent conductor layer. The term superstrate is used since the glass faces the incoming sunlight. Reprinted from Luque, A., and Hegedus, S., Handbook of Photovoltaic Science and Engineering, ISBN: 978-0-471-49196-5. Copyright (2003) with permission from John Wiley & Sons*

CdTe in conjunction with n-type In_2O_3:Sn, ZnO, SnO_2 or CdS have been extensively investigated. The most effective thin-film device is experimentally found to be the p-CdTe/n-CdS structure.

CdTe and CIGS solar cells can achieve well over 10% efficiency without the need for tandem structures, which gives them a cost advantage over silicon; however, there is considerable controversy surrounding the long-term viability of these alternative materials systems. Cd is toxic and is used in CdTe cells as well as CIGS devices, which also require a CdS layer. An increasing number of countries are restricting the use of Cd. In addition, the world supply of elements such as Te, In and Ga is not currently sufficient to enable unrestricted growth in manufacturing volume of these thin-film solar cells.

In spite of these challenges CdTe solar cells broke a long-standing cost barrier in 2008–2009 and can now be produced for under US$1 per watt. CdTe is attractive for large-scale solar farms where lower efficiency solar cells are adequate, potential issues of toxicity can be well managed, and low cost is essential.

4.13   High-Efficiency Multijunction Solar Cells

An important area of active development of solar cells involves increasing efficiency. The motivation for this comes from several factors. For space applications, weight is crucial. For terrestrial applications, the opportunity to concentrate sunlight onto smaller solar cells is attractive (see Section 4.14). A premium price *per watt* can therefore be charged for higher efficiency solar cells. In addition cost reductions in substrate and packaging materials and transportation and installation are attractive motivators.

The achievement of 30–40% conversion efficiency in single-crystal multijunction solar cells has been an important milestone in solar cell development (Figure 4.16). Of importance for space applications, these high-efficiency cells are also useful in spite of their high cost for terrestrial *concentrating solar cell* installations in which sunlight is concentrated using mirrors or lenses to yield more power per unit area of solar cell. Concentration factors of several hundred are achievable, although *tracking* or angular movement of the solar cell is required to follow the sun's position. See Section 4.14.

The most successful tandem single-crystal solar cell uses germanium as a substrate. Germanium has a convenient lattice constant since it matches the lattice constant of GaAs (see Appendix 2). In addition, Ge has a small bandgap allowing it to absorb longer wavelengths of light than silicon. By growing thin-film epitaxial layers of III-V semiconductors on the Ge, a tandem cell with two or three junctions can be formed that absorbs effectively over a wide wavelength range of solar radiation.

The structure of a triple junction solar cell is shown in Figure 4.29. The GaAs thin-film cell is epitaxially grown on the Ge substrate. A third semiconductor, $Ga_{1-x}In_xP$, is grown epitaxially on the GaAs. From Figure 1.33a it can be appreciated that the lattice constant of $Ga_{1-x}In_xP$ can be set to match GaAs provided the appropriate value of x is chosen. This value of x is close to 0.5 and the resulting bandgap is about 1.85 eV. The triple junction solar cell therefore has three bandgaps of 0.7 eV (Ge), 1.43 eV(GaAs) and 1.95 eV($Ga_{1-x}In_xP$). A useful way to visualize the principle behind multijunction solar cells is shown in Figure 4.30.

Future high-efficiency solar cells based on the multijunction approach will require lower cost materials in both bulk form and thin-film form. The use of germanium wafers is
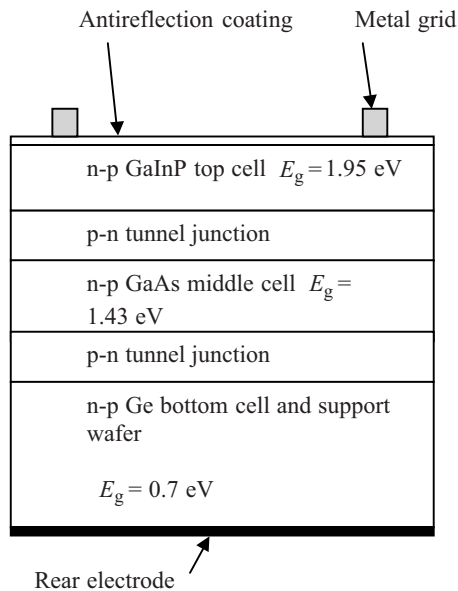
Figure 4.29   *Structure of triple junction solar cell showing germanium wafer upon which two larger bandgap semiconductors are deposited epitaxially. The total output voltage is obtained between the metal grid and the rear electrode*
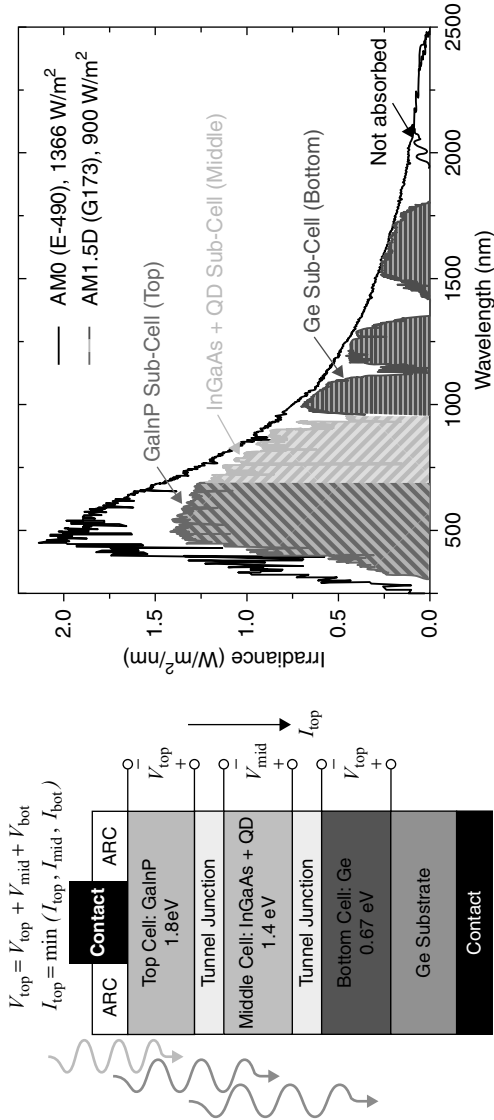
Figure 4.30 *Multijunction solar cell showing the portions of the spectrum that are covered by each material. Note that the infrared wavelengths above 2000 nm are not absorbed. The GaAs junction is modified and further optimized by the addition of a small indium component. Note that the three cell voltages add together, however the cell current is controlled by the minimum current that is supplied by each of the cells. This means that current matching between the multiple cells in a multijunction cell is important. Data provided courtesy of Spectrolab Inc. Copyright (2011) Spectrolab*

not cost effective for non-concentrating terrestrial applications. The multijunction solar cell is an area of active development driven by the promise of still higher efficiencies and lower production cost per watt. Theoretical models show that efficiencies of over 60% are possible in a multijunction cell provided the number of junctions is increased to five to eight junctions in tandem. The potential exists eventually to achieve cost-effective high-efficiency multijunction solar cells for terrestrial use without the need for optical concentration however as noted in Figure 4.30, current matching between cells is critical to prevent a weak link in the series circuit of cells.

4.14   Concentrating Solar Systems

The high cost of multijunction solar cells can be very effectively offset by *concentrator solar technology*, which concentrates sunlight onto solar cells using either reflective optics in the form of curved mirrors, or refractive optics in the form of traditional lenses or Fresnel lenses. Concentration of sunlight intensity by factors of several hundred or even a thousand are possible. This approach is attractive but its application is generally limited to large-scale solar farms due to the difficulties associated with it:

(a) The optical system must be aimed at the sun to maintain maximum solar power on the solar cell at all times of the day and over all seasons. This requires *tracking* devices that move the optical system mechanically. Single-axis trackers only move in one axis and can follow the sun over the course of one day; however, it is more optimal to use dual-axis tracking to allow for seasonal variations of sunlight. The higher the solar cell gain the more precise the tracking needs to be since the focus of the sunlight will fall onto a smaller area.

(b) The high intensity will cause a temperature increase at the solar cell due to light that is absorbed but does not directly contribute to electrical power. Since even the best commercial multijunction solar cells have efficiency values of under 40%, more heat energy than electrical energy is produced and cooling is generally required. For high concentration factors of several hundred, liquid cooling may be used, although air cooling is simpler and generally less costly.

(c) The efficiency of crystalline silicon solar cells decreases as temperature rises, which will offset some of the benefits of solar concentration including the efficiency gain noted in Figure 4.15 (see Example 4.2 and Problem 4.7). This, although not as significant as in silicon, is still an issue for wider bandgap III-V semiconductor solar cells and multijunction cells.

(d) On cloudy days, non-concentrating solar cells will still generate a sizeable fraction of their full-sun power output; however, concentrator systems will not function well due to the diffuse nature of the sunlight.

For these reasons the deployment of non-tracking solar cells is often preferred. Both their better appearance and their reliability and simplicity are important advantages of non-tracking cells, especially in urban areas and rooftop systems. Ultimately a market exists for both tracking solar cell systems and non-tracking systems. Some of these aspects are further explored in Problem 4.9.

4.15   Summary

4.1. A solar cell consists of a p-n junction in which sunlight generates electron-hole pairs that are collected by the external circuit and provide electrical power. Carrier flow occurs by drift across the depletion region in the solar cell and operation is in the fourth quadrant of an *I–V* diagram. The potential barrier in a solar cell p-n junction at the operating point should be as small as possible to maximize the output voltage of a solar cell, but must be large enough to maintain a suitable reverse current.

4.2. Light absorption in a solar cell is sensitive to the type of bandgap. Indirect gap semiconductors have weaker absorption than direct-gap semiconductors over a range of photon energies. The absorption in indirect gap material requires the participation of phonons and thicker cells are required.

4.3. The solar spectrum is based on the blackbody radiation spectrum at a temperature corresponding to the surface of the sun at 5250°C. This spectrum is affected by atmospheric absorption due principally to oxygen, water vapour and carbon dioxide.

4.4. A crystalline silicon solar cell consists of a thin n-type front layer and a thick p-type layer. A front electrode metal grid, back contact metallization and an antireflection layer complete the device. Sunlight that is absorbed within the depletion region or within a diffusion length on either side of the depletion region has the best chance of generating power in an external circuit. Short circuit current, open circuit voltage and fill factor are useful characterization parameters. Show that Equation 4.3a is obtained from Equation 4.2 by applying the appropriate boundary conditions.

4.5. In 'thin' solar cells diffusion lengths exceed the thickness of the cell. Surface recombination velocity is an important parameter that allows us to quantify the chances of carriers recombining at surfaces instead of through the external circuit.

4.6. The generation of usable current as a function of depth in a solar cell may be determined by assuming light is absorbed at a specific depth and then determining the collection of holes and electrons from this depth.

4.7. The limiting solar cell efficiency is dependent on the bandgap of the solar cell. Limitations in efficiency arise from photons that are too low in energy to be absorbed or photons with too much energy that must thermalize before they can be collected. In addition the operating voltage is lower than the bandgap energy, which represents a further energy loss. Silicon has achieved a conversion efficiency of approximately 25%.

4.8. Silicon for silicon solar cells starts with metallurgical grade material that is most commonly refined using the Siemens process. The resulting solar grade materials may be cast into multicrystalline ingots or grown into boules by the Czochralski method. Wire sawing is used to cut wafers. Single-crystal silicon is a few percent more efficient that multicrystalline silicon.

4.9. Solar cell wafer processing includes texturing, back and front surface doping and contacting, and the deposition of an antireflection layer. Screen printing offers a low-cost approach to processing.

4.10. Silicon ribbon technology is an alternative to wafer dicing for solar cells. Silicon is pulled directly from the melt as a liquid held between two wires by surface tension.

4.11. Thin-film silicon amorphous solar cells use a random silicon structure with no long-range order but predominantly tetrahedral bonding. Dangling bonds are terminated

by hydrogen. The bandgap of amorphous silicon is about 1.75 eV and it behaves more like a direct-gap semiconductor. The p-i-n structure is used to fabricate solar cells due to short carrier lifetimes in doped material. Tandem and triple junction designs can increase conversion efficiency by up to approximately 10%.

4.12. Thin-film solar cells based on CdTe and CIGS achieve over 10% efficiency without the need for multijunction architectures. Both CdTe and CIGS are direct-gap semiconductors. CdTe solar cells are the first to break the US$1 per watt cost barrier.

4.13. Multijunction solar cells have achieved over 30% conversion efficiency. The best devices are grown by epitaxial layer growth of III-V semiconductors lattice matched to a germanium substrate. The solar spectrum is more efficiently utilized than in single-junction solar cells and the maximum theoretical efficiency is above 60%.

4.14. Concentrating solar systems allow for sunlight to be concentrated onto smaller, high-efficiency solar cells. This can justify the cost of multijunction solar cells; however, tracking systems are required to maintain optimum system efficiency, and sunlight diffused by clouds does not concentrate effectively.

Suggestions for Further Reading

Green MA. *Solar Cells, Operating Principles, Technology and System Applications*. University of New South Wales, 1992.

Luque A and Hegedus S (eds). *Handbook of Photovoltaic Science and Engineering*. Wiley, 2003. Streetman BG and Banerjee SK. *Solid State Electronic Devices*, 6th edn. Prentice Hall, 2006.

Problems

4.1 Write a short two-to-three page qualitative description in your own words of the processes involved in producing electric power from a p-n junction solar cell. Include diagrams as necessary. You should refer specifically to the following:
optical absorption
electron-hole pair production
carrier drift/diffusion
recombination
diffusion length
open circuit voltage
short circuit current

4.2 Solar radiation is fundamentally determined by *blackbody radiation*. Blackbody radiation produces a radiation spectrum described by *Planck's law*. Planck's law describes the *radiance* as a function of frequency ν of electromagnetic radiation emitted from a non-reflective surface of a body in thermodynamic equilibrium at temperature T. As a function of ν in cycles per second and T in kelvins, Planck's law is

$$I(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{\exp\left(\dfrac{h\nu}{kT}\right) - 1}$$

and represents the emitted power in watts per unit area of an emitting surface at temperature T per unit solid angle per unit frequency. It provides what is referred to as a *specific radiative intensity*.

(a) Using a computer, plot $I(\nu, T)$ as a function of ν on a single set of coordinates for the following temperatures: $T = 1500$ K, $T = 2500$ K, $T = 3500$ K, $T = 4500$ K and $T = 5500$ K. For each plot estimate the perceived colour of the observed emission and find the wavelength at which $I(\nu)$ is a maximum. At what wavelength is sunlight a maximum?

(b) For each temperature, find the fraction of total radiated power that falls in the visible light range between 400 and 700 nm. *Hint:* Use the area under the curve of $I(\nu, T)$.

(c) The CIE diagram of Figure 3.12 shows the locus of colour coordinates x and y due to blackbody radiation. Show the point on this locus corresponding to a tungsten filament lamp operating at 3500 K. How much below the melting point of tungsten is this? Also show the point on the locus corresponding to the light emitted by the sun.

4.3 Antireflection layers applied to silicon solar cells function in a manner similar to those used on camera lenses and eyeglass lenses. If light is incident normal to a first interface between two transparent optical media with different indices of refraction, reflection will take place. If the reflection at a second interface is arranged to cancel out the first reflection by ensuring that both the magnitude and the phase of the reflections will cancel out the first reflection then overall reflection is reduced or

eliminated. The reflection coefficient R is

$$R = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2$$

A simple example of this is the interface between air and glass shown below:
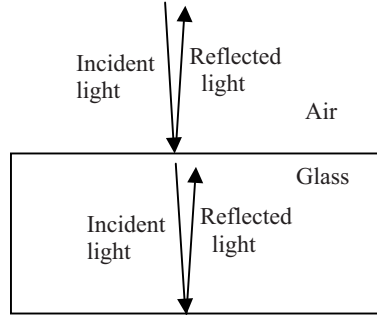
Air has index of refraction $n_1 \cong 1.0$ and glass has index of refraction $n_2 \cong 1.5$. Therefore,

$$R = \left(\frac{n_2 - n_1}{n_2 + n_1}\right)^2 = \left(\frac{1.5 - 1.0}{1.5 + 1.0}\right)^2 = 0.04 \text{ or } 4\%$$

Note that the same reflectivity is obtained from the air–glass interface whether the light is incident from air onto the air–glass interface or from the glass onto the glass–air interface.

An antireflection layer having index of refraction n_2 as shown below creates two interfaces, one between media 1 and 2, and one between media 2 and 3:
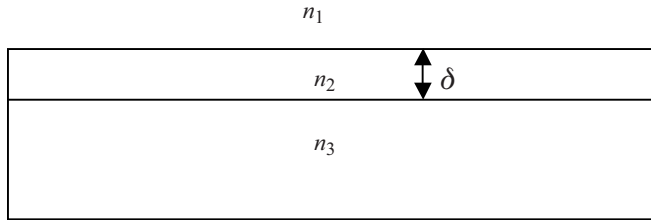
(a) Regard n_1 and n_3 as given, but n_2 as an adjustable parameter. Find the value of n_2 in terms of n_1 and n_3 that will cause the reflection from the interface between 1 and 2 to be the same in magnitude as the reflection from the interface between 2 and 3.

(b) If medium 1 is air and medium 3 is glass, find the numerical value for n_2 that satisfies the criterion of (a).

(c) In order for the reflections to cancel out, the phases of the light waves leaving the interfaces must also be arranged to be 180 degrees out of phase with each

other. Show that the smallest possible thickness δ of medium 2 is $\frac{\lambda}{4}$ where λ is the wavelength of the light in medium 2.

(d) For the conditions in (b) find the smallest possible thickness of medium 2 to allow for cancellation of reflection for light having wavelength 600 nm. *Hint:* Wavelength in a medium having index of refraction n is reduced relative to its wavelength in vacuum by a factor of n.

(e) Assuming that silicon is a transparent medium, and using information in the appendices, find both the thickness and index of refraction of an antireflection layer that could be deposited on silicon if it is situated in air.

(f) Tantalum oxide is a popular antireflection material for silicon solar cells. Look up the refractive index of tantalum oxide and compare with your answer to (e).

(g) Since silicon solar cells are generally bonded to a glass front sheet with $n \cong 1.5$, repeat steps (e) and (f) if the air is replaced with glass. In which case (air or glass) does tantalum oxide best match the calculated refractive index? You may neglect the front reflection off the glass sheet at the air interface.

(h) Compare the expected efficiency of a silicon solar cell with an antireflection layer to that without an antireflection layer. *Hint:* Light that is not reflected will be transmitted into the silicon. Assume a wavelength of 550 nm.

(i) Silicon solar cells with an antireflection coating generally have a deep-blue colour. Explain this based on your understanding of antireflection layers.

4.4 Show that Equation 4.2 is the solution to Equation 4.1 by substituting it into the differential equation. Also show that Equation 4.3 is obtained from Equation 4.2 by applying the appropriate boundary conditions.

4.5 Using the expressions for diffusion current,

$$J_n(x)_{\text{diffusion}} = qD_n \frac{dn(x)}{dx}$$

and

$$J_p(x)_{\text{diffusion}} = -qD_p \frac{dp(x)}{dx}$$

show that Equations 4.4a and 4.4b are obtained from Equations 4.3a and 4.3b.

4.6 The optical generation rate inside a silicon solar cell at room temperature in a certain flux of sunlight is assumed to be a constant value $G = 5 \times 10^{21} \text{EHP cm}^{-3} \text{ s}^{-1}$. If $W = 1.0$ μm, the electron and hole lifetimes are 2×10^{-6} s, $I_0 = 1 \times 10^{-9}$ A and the cell area is 100 cm^2, calculate

(a) The optically generated current I_L.

(b) The short circuit current I_{SC}.

(c) The open circuit voltage V_{OC}.

(d) The maximum output power if the fill factor is 0.8.

4.7 A silicon solar cell of area 100 cm^2 has doping of $N_a = 5 \times 10^{18} \text{cm}^{-3}$ and $N_d = 5 \times 10^{16}$ cm^{-3}. If the carrier lifetimes are 5×10^{-6} s and the optical generation rate inside the silicon solar cell is assumed to be a constant value of $G = 7 \times 10^{21}$ EHP cm^{-3} s^{-1}, calculate

(a) The optically generated current I_L. State and justify any assumptions you used to obtain your solution.

(b) The short circuit current I_{SC}.

(c) The open circuit voltage V_{OC} at room temperature.

(d) The open circuit voltage at 100°C and at −50°C.

(e) The maximum output power if the fill factor is 0.8 at both room temperature and at 100°C.

4.8 Section 4.6 introduced a method of calculating the contribution to solar cell current in a thick solar cell from electron-hole pair generation at a specific depth. This is instructive but not physically realizable since light will exist at more than one depth. The approach used in Section 4.4 was to assume that light is uniform in intensity inside the solar cell both within the depletion layer and within a diffusion length on either side of the depletion layer. This is also not physically realizable since light is being absorbed as it goes deeper into the solar cell.

(a) Explain how an improved model could build on the method of Section 4.6 to model a solar cell more accurately. Your model should begin by obtaining the intensity of sunlight as a function of depth. You do not need to fully formalize all the equations in your model but make sure you demonstrate an understanding of the approach.

(b) Can you extend the model further to allow it to handle sunlight with the full solar spectrum? What additional considerations are needed in your model?

4.9 Concentrating solar cells require an optical means of concentrating sunlight onto solar cells.

(a) Research concentrator solar cells and find three designs that have been used for concentration of sunlight. Explain their optical design using a diagram. You should make sure you have at least one reflective design and one refractive design, although there are designs that use both reflection and refraction.

(b) The concentrator designs allow for the use of high-efficiency multijunction cells. If concentration by a factor of 200 is achievable and multijunction solar cells with efficiency of 30% are used in place of silicon cells having 15% efficiency, how many cloudy days per year would cancel out the benefit of the efficiency improvement in terms of energy produced over one year? Assume a cloudy day reduces the power output from a non-tracking cell by a factor of three, and assume that the concentration factor is eliminated by the clouds. Make and justify any additional assumptions you use.

4.10 One solar panel consists of 80 15-cm diameter silicon wafer-based cells laminated onto a glass plate.

(a) Calculate the maximum power output from the panel. Assume the silicon cells are 18% efficient.

(b) Sunlight, over the course of one day, is modelled as a light source that rises and sets and at noon the sunlight is incident normal to the cell whereupon it provides $1000\,\mathrm{W\,m^{-2}}$ of optical power. Calculate the total energy output in units of watt-hours over the course of one day. Make and justify any other assumptions needed.

(c) A typical family home is frequently assumed to require approximately $3000\,\mathrm{W}$ of solar panel capacity to satisfy its energy needs. Justify this based on your answer to (b) by considering the energy requirements of a home including heating, hot water and lighting. How many solar panels would be required? Make a table to show energy consumption for a range of uses in kWh (kilowatt-hours).

4.11 One of the most important thin-film solar technologies is based on microcrystalline silicon. Efficiency levels of approximately 9% are routinely achieved in production. Mass-production technology for stacked triple-junction thin-film solar cells has been achieved by adding another amorphous layer to a conventional two-active-layer structure (amorphous silicon plus microcrystalline silicon) resulting in a triple-junction structure with amorphous silicon (two active layers) and microcrystalline silicon (single active layer). This architecture achieves a cell conversion efficiency ranging from 11% to 13% and module conversion efficiency from 8.6% to 10%. Mass production began in 2007. Creating two amorphous silicon active layers increased voltage levels, and structuring the cell to have three active layers in combination with microcrystalline silicon decreases light-induced degradation (drop in conversion efficiency).

    (a) Sketch your understanding of the device structure. What would you estimate the open circuit voltage to be?

    (b) Look up the appropriate information on the web and summarize the performance specifications of commercially available products. Recommended key words: triple-junction thin-film silicon, Sharp corporation (Japan).

    (c) What advantages does this technology have compared to CdTe thin-film solar cells?

    (d) What disadvantages does it have compared to CdTe thin-film solar cells?

4.12 Cadmium telluride thin-film modules achieved an installed capacity of over 2 GW in 2010. They require only 2% of the semiconductor content of crystalline silicon solar cells. Read literature available from the thin-film CdTe industry. Summarize the available modules and their efficiency and power output specifications. Recommended key words: First Solar Corporation (USA), CdTe thin-film solar cells.

    (a) How are toxicity and potential environmental issues addressed by the manufacturer?

    (b) List and explain two other positive attributes of CdTe thin-film solar panels as presented by CdTe PV manufacturers.

    (c) The installation of solar farms is the biggest use of CdTe technology. From the list of the ten largest photovoltaic solar farms below, as of December 2010, determine which projects use CdTe technology. For the solar farms not using CdTe, find the solar cell technology in use.

    (d) The average output of these ten largest solar farms will be greatly exceeded by solar farms currently either under construction or planned for construction. Find information on these near-term projects and compare the capacities with the above list.

    (e) To put the combined power output of PV solar farms in context, look up the world electricity generating supply and determine what fraction of electricity production is in the form of PV solar.

4.13 The prospect exists for obtaining over 50% conversion efficiency in a multijunction solar cell with more than three solar cells in tandem. If both the II-VI and the III-V solar cell materials are used, a set of five junctions using direct gap semiconductors is available; see Figure 4.31.

    (a) Explain why an increasing number of junctions in tandem could, in principle, allow for a higher overall efficiency for a photovoltaic solar cell operating with illumination from the sun.

Table 4.1   *Ten largest solar photovoltaic power stations at December 2010*

PV power station	Country	Nominal power (MW$_p$)
Finsterwalde Solar Park	Germany	80.7
Sarnia Photovoltaic Power Plant	Canada	80
Rovigo Photovoltaic Power Plant	Italy	70
Olmedilla Photovoltaic Park	Spain	60
Strasskirchen Solar Park	Germany	54
Lieberose Photovoltaic Park	Germany	53
Puertollano Photovoltaic Park	Spain	47.6
Moura photovoltaic power station	Portugal	46
Kothen Solar Park	Germany	45
Waldpolenz Solar Park	Germany	40

*The ratio of the total energy a plant produces during a period of time to the energy produced at full capacity.
MWp, megawatts peak; GWh, gigawatt hours.
Adapted from Wikipedia (see http://en.wikipedia.org/wiki/List_of_photovoltaic_power_stations).

  (b) From the figure, describe the possible structure of a solar cell with five tandem junctions. What substrate could be used to provide a lattice constant suitable for all the layers? Is such a substrate available?

 4.14 The single-crystalline silicon solar cell may be further optimized by a back electrode structure in which both the anode and cathode contacts are attached to the rear of the cell in closely spaced stripes. This cell architecture has the advantage that the front of the silicon solar cell is completely available for sunlight absorption and no power is blocked by front electrodes. Since the electrodes are all on the back of the
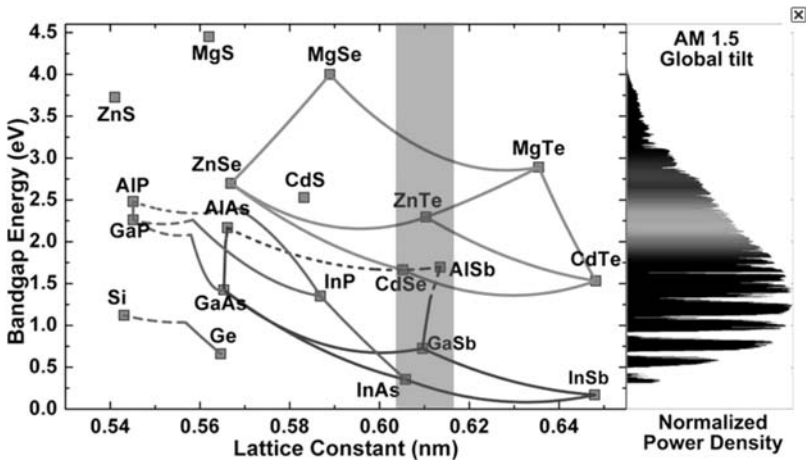
Figure 4.31   *Reprinted with permission from Zhang, Y. H., and Ding, D., Arizona State University*

wafer, carriers must flow laterally inside the wafer and electrons and holes will drift to adjacent electrode stripes on the back of the wafer. The overall efficiency of these cells is a remarkable 24% in production quantities. Find additional information on this type of solar cell. Prepare a two-to-three-page summary of the cell structure and the important distinctions between these rear electrode cells compared to the traditional front/rear electrode cells.

Hint: Use the following search terms: SunPower rear electrode solar cell.

5

Light Emitting Diodes

Objectives

1. Understand basic LED device operation and LED device structures based on the p-n junction.
2. Understand the LED emission spectrum and emission linewidth based on direct gap recombination.
3. Review the sources of non-radiative recombination that must be recognized and minimized in effective LED design.
4. Introduce and model optical outcoupling concepts that determine the efficiency with which light can emerge from LED devices.
5. Describe GaAsP LED technology and performance levels.

6. Describe AlGaAs LED technology and introduce the double heterostructure design and performance levels.
7. Describe AlGaInP LED technology and performance levels.
8. Describe GaInN LED technology and performance levels.
9. Review approaches to improve outcoupling from LED structures and review LED structures designed for high-power applications.
10. Introduce the principles of down-conversion of LED light to create white-emitting LEDs.

5.1   Introduction

The inorganic light emitting diode (LED) has shown remarkable development since the early 1960s when the first practical red emitting devices were introduced to the market. Lighting during the twentieth century was dominated by fluorescent, incandescent and gas discharge lamps, and the LED was regarded as a low-power and relatively low-efficiency light source that had the advantage of long life and low voltage operation. This gave the LED a place in early portable electronics such as calculator and watch displays in the 1960s; however, these applications were replaced by low-power liquid crystal display technology. LEDs grew slowly in importance as LED-based indicator lights and small LED numeric and alphanumeric displays provided improved viewability compared to liquid crystal displays for consumer and industrial applications from clock radios to industrial ovens. The prevailing advantages of their bright and attractive appearance and very long life enabled market penetration and increasingly lower production costs.

The development of LED semiconductor materials and device structures has been driven since the 1960s to achieve a wider range of colours as well as increases in efficiency. The earliest generations of LEDs were red and infrared, and the ongoing challenge has been to increase the bandgap of the semiconductors available for LED fabrication, and to develop effective manufacturing methods for these materials. The availability of both n-type and p-type LED material is essential and doping becomes increasingly difficult as bandgap increases.

During the decades following the 1960s, bandgaps gradually increased in stages allowing yellow and yellow-green LEDs to be manufactured; however, it was not until the 1990s that bright-blue and saturated green LEDs were achieved. By this time red and yellow LEDs had matured, and improvements in design and device structure enabled high efficiency levels with sufficient brightness for LEDs to be seen outdoors. These device structure developments in red and yellow LEDs were rapidly transferred to the new green and blue LED materials to yield the full colour range of red, green and blue LEDs. New, very large markets for LEDs now emerged in outdoor video signage, traffic lights and external automobile lights.

Another development during the 1990s was achieved by combining a blue LED with a yellow-emitting phosphor material to create a bright white light emission for illumination purposes. In applications where long life and low power consumption are important enough to offset the higher cost compared to incandescent or fluorescent lighting, white LED lighting is gaining prominence. Sectors that must maintain public places where reliable

illumination is required over extended time periods include transportation and retail sectors. As white LED costs drop they are being increasingly introduced to consumers as alternative lighting products for the home due to their flexibility and small size combined with their low power consumption and long life.

Another large market has emerged recently in the display industry. As liquid crystal displays (LCDs) grow in size and importance LEDs have developed an increasing market share of the LCD backlighting market. Full-colour LCDs require bright trichromatic (red-green-blue) backlighting. In portable products such as cellphones, LEDs routinely backlight LCD displays and were adopted early due to their low-voltage operation; however, larger LCD displays in laptop computers used fluorescent lamp backlighting. This is now changing rapidly, and LED backlighting is emerging as the best backlighting approach for laptop and desktop computers as well as LCD televisions due to their higher efficiency compared with fluorescent lamps. LEDs also have the advantage that they do not contain the mercury that is required for fluorescent lamp operation, which is an important advantage in terms of product disposal and recycling.

Today, efficient LED emission, not only in the visible wavelength range, but also in the UV and IR ranges is available. It should be noted that UV and IR emitting devices are still commonly called LEDs. These devices find application in IR remote controls and as UV light sources for industrial processes such as for promoting chemical reactions that are triggered by a UV light source.

Inorganic LEDs are p-n junction devices, as are PV solar cells; however, they are at opposite ends of the size range of semiconductor devices. LEDs typically are diced into individual chips or die that have junction areas of $\cong 0.1\,mm^2$ whereas solar cells have corresponding areas of $\cong 100\,cm^2$, which is a factor of 10^5 larger. Larger area LEDs up to $\cong 1\,cm^2$ in area do exist for high-power applications, but these devices require effective heat-sinking to maintain a suitable operating temperature.

Fundamental aspects of LED device physics and LED device design are covered in this chapter, and the development of both the semiconductor materials and device design improvements are presented. Sections 5.6 to 5.10 review materials and devices as they developed since the 1960s.

5.2 LED Operation and Device Structures

Figure 5.1 shows a historical plot of the efficiencies and colours achieved by LEDs over the past half century. Of significance is the achievement of LED efficiency values that exceed other traditional lamp technologies such as incandescent lamps and fluorescent lamps.

The simplest LED p-n junction is illustrated in Figure 5.2. Light is emitted when electrons and holes that flow as majority carriers are injected across the junction and recombine at or near the junction causing radiative recombination to take place. This is a diffusion-driven process near the junction as opposed to the drift mechanism that underlies solar cell operation. Radiative recombination is favoured in direct-gap semiconductors and as a result the materials chosen for most practical LEDs are direct gap. There are exceptions to this provided that some mechanism exists to overcome the lack of radiative recombination in indirect gap material. This will be further explained in Section 5.4.

LED device structures vary; however, the most common structure of a small LED showing both the packaging as well as the semiconductor die is illustrated in Figure 5.3. Since light
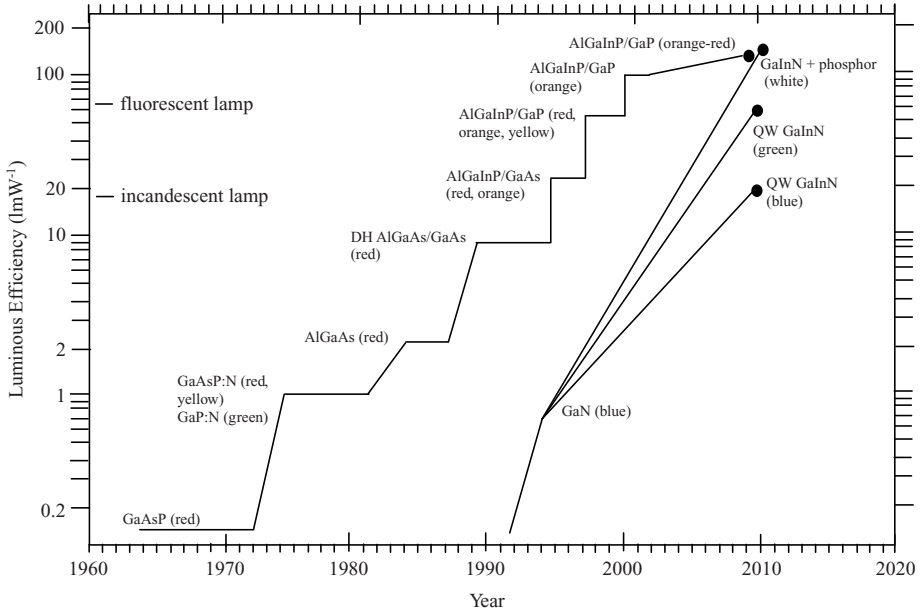
Figure 5.1   *Efficiency of LEDs achieved as a function of time since the 1960s. The photometric quantity lm W^{-1} (lumens per watt) was introduced in Chapter 3*

is emitted from the sides as well as the front of the die, a reflective cup is used to allow the light to propagate forwards and through a moulded epoxy polymer lens.

A more detailed view of the die is shown in Figure 5.4. This shows a single crystal n-type substrate used to support the active layers, which will be discussed in detail in this chapter. A thick p-type layer above the active layers spreads the current from the top contact over
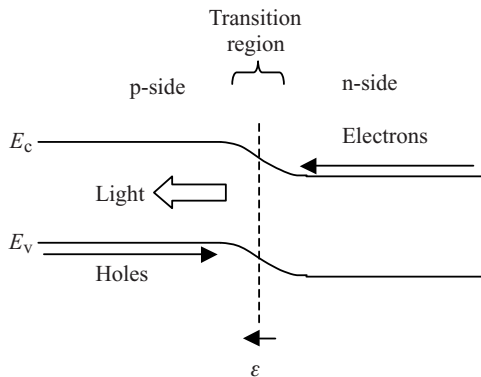
Figure 5.2   *Forward-biased LED p-n junction. When one electron and one hole recombine near the junction one photon of light may be emitted. The achievement of high efficiency requires that there is a good chance that recombination events are radiative and that the generated photons are not reabsorbed or trapped in the device*
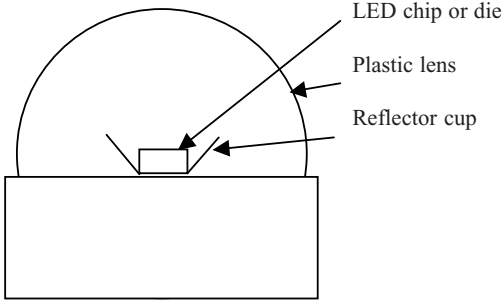
Figure 5.3   *LED packaging includes a transparent lens, which is usually made from an epoxy material, and a reflector cup into which the LED die is mounted. The radiation pattern is determined by the combination of the die emission pattern, the reflector cup design, and the shape and refractive index of the polymer lens. Reproduced by permission of Avago Technologies*
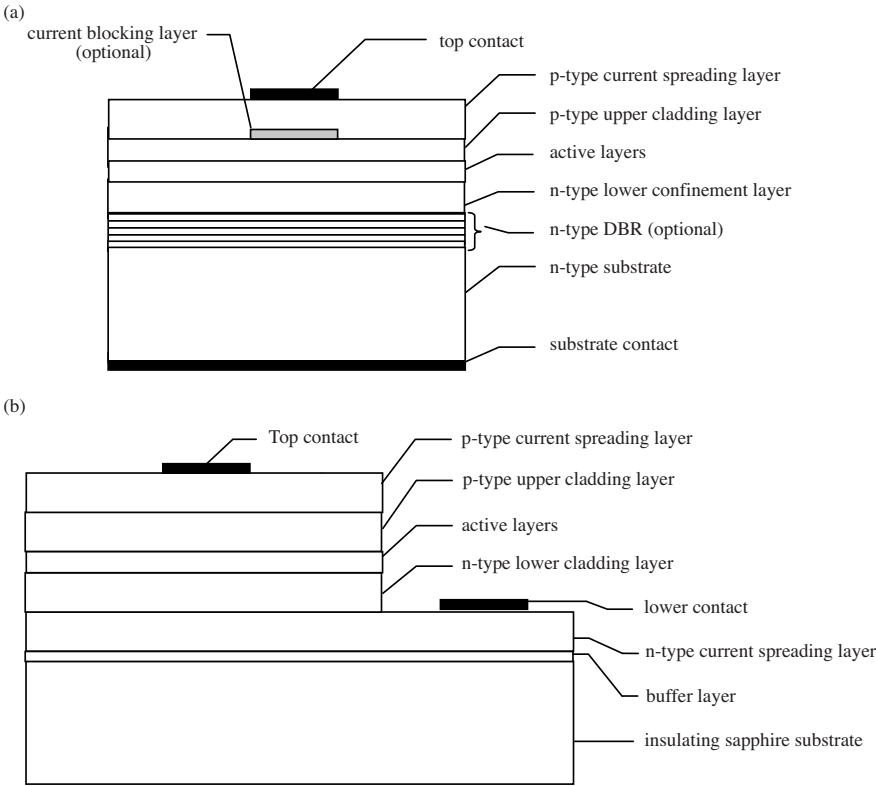
Figure 5.4   *LED die consists of a single crystal substrate on which a series of epitaxial layers is grown forming the active layers. (a) For conductive substrates a current spreading layer and the top and bottom contacts are shown as well as an n-type current blocking layer discussed in section 5.4. The notations DH and DBR refer to Double Heterostructure and Distributed Bragg Reflector, which will be discussed in sections 5.8 and 5.11 respectively. (b) For LEDs grown on insulating substrates such as sapphire (see section 5.10) the second contact is made via a buried current spreading layer as shown*
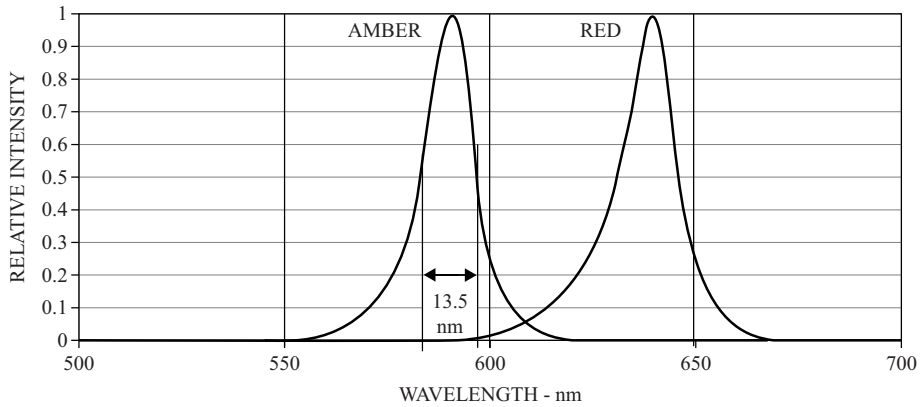
Figure 5.5   *Emission spectra of AlGaInP LEDs. The linewidth of the amber LED is measured as the full width at half maximum as shown and is 13.5 nm. Reproduced by permission of Avago Technologies*

the junction area. The top contact is made as small as possible to avoid blocking generated light leaving through the top of the die.

5.3   Emission Spectrum

Consider the direct-gap semiconductor having approximately parabolic conduction and valence bands near the bottom and top of the bands respectively. The photon emission rate $R(E)$ of a LED was derived in Chapter 3, and the result from Equation 3.21 is

$$R(E) = R(h\nu) \propto (E - E_{\mathrm{g}})^{\frac{1}{2}} \exp\left[-\frac{E}{kT}\right] \qquad (5.1)$$

From this the intensity profile of a LED may be plotted as a function of wavelength λ where $E = \frac{hc}{\lambda}$. The result was shown graphically as a function of energy in Figure 3.10. The same data can be replotted on an intensity versus wavelength graph (Figure 5.5), which shows measured emission spectra from commercially available red and amber LEDs.

Example 5.1

Obtain the full width at half maximum (FWHM) for the amber LED of Figure 5.3 and compare with the theoretical value.

Solution

From Section 3.8, the full width at half maximum (FWHM) for a LED is found to be $\Delta E = 1.8kT$ at room temperature (see Figure 3.10). For the amber LED of Figure 5.5 the graph gives us $\Delta \lambda = 13.5$ nm.
   At the lower FWHM point, reading the graph, $\lambda_{\mathrm{lower}} = 584$ nm.
   At the upper FWHM point, reading the graph, $\lambda_{\mathrm{upper}} = 597$ nm.

Now,

$$E_{\text{lower}} = \frac{hc}{\lambda_{\text{lower}}} = \frac{(6.63 \times 10^{-34} \text{ Js})(3 \times 10^8 \text{ m s}^{-1})}{584 \times 10^{-9} \text{ m}}$$

$$= (3.40 \times 10^{-19} \text{ J}) \left(\frac{1}{1.6 \times 10^{-19}} \frac{\text{eV}}{\text{J}} \right) = 2.13 \text{ eV}$$

and

$$E_{\text{upper}} = \frac{hc}{\lambda_{\text{upper}}} = \frac{(6.63 \times 10^{-34} \text{ Js})(3 \times 10^8 \text{ m s}^{-1})}{597 \times 10^{-9} \text{ m}}$$

$$= (3.33 \times 10^{-19} \text{ J}) \left(\frac{1}{1.6 \times 10^{-19}} \frac{\text{eV}}{\text{J}} \right) = 2.08 \text{ eV}$$

Hence, $\Delta E = 2.13 - 2.08 = 0.05$ eV.

From Section 3.8, at 300 K:

Emission linewidth $= 1.8 \, kT = 1.8 \times 0.026 = 0.047$ eV in reasonable agreement with the measured result.

Other broadening mechanisms exist for LED emission including inhomogeneous broadening in which non-uniformities in LED material modify the semiconductor locally. This is particularly prevalent for alloy semiconductors where small deviations in both composition and defects are possible. Emission intensity, peak emission wavelength and linewidth typically vary from die to die even if they are from the same semiconductor wafer.

For this reason visible emitting LEDs intended for applications where device-to-device uniformity is important are *binned*, or sorted into batches based on measured values of emission colour and luminous intensity. Units used to specify visible light emitting LEDs include luminous intensity and colour coordinates which were introduced in Section 3.9.

Ultimate LED efficiency is limited by two factors, namely:

(a) *non-radiative recombination* in which electron-hole pairs recombine but do not generate a photon;
(b) limited *optical outcoupling* in which not all the generated photons leave the device.

The issue of optical outcoupling plays a key role in LED performance. Generated photons can be reabsorbed in LEDs, and they may be reflected back into the device by surface or interface reflections. A number of device structures have been developed to optimize outcoupling. These will be discussed in this chapter.

5.4　Non-Radiative Recombination

Non-radiative recombination can occur through a variety of mechanisms. These include traps and surface states, which were discussed in Section 1.19. Additional traps can occur due to native defects such as interstitials and vacancies, and in compound semiconductors *anti-sites* can result in energy levels within the forbidden gap. Anti-sites refer to the

substitution of a cation in the place of an anion or an anion in the place of a cation in a compound III-V or II-VI semiconductor. Particularly if the levels are near mid-gap they become traps and operate effectively to cause non-radiative recombination.

In addition, as the operating temperature of the LED rises, it promotes electrons to higher energy states in the conduction band, which has the effect of giving a nominally direct-gap semiconductor some indirect-gap character. This can occur since electrons can be thermally promoted to occupy a secondary valley in the conduction band of direct gap semiconductors. These electrons are then forced to recombine indirectly. See band diagrams in Figure 1.16 for a variety of semiconductors, including GaAs, that have a primary direct-gap conduction band valley as well as a secondary indirect gap valley. The smaller the energy barrier between the primary and secondary valleys the easier it becomes to thermally excite electrons between them.

Another important mechanism for non-radiative recombination involves *Auger recombination*. Here an electron near the bottom of the conduction band can be promoted to a high-energy state within the conduction band due to the recombination of a hole-electron pair. Rather than a photon, the result is lattice heating as the high-energy electron subsequently thermalizes. A similar mechanism involving a hole is also possible.

Auger processes become important especially at high rates of recombination. This is because either two electrons or two holes are required for this process and the chance of having the carrier pair available to take part in the Auger process is therefore proportional to the square of the carrier concentration.

Surface recombination is also an important process. This was discussed in Section 1.19. Due to the high electron mobility in materials such as GaAs, surface recombination velocities are high. It is therefore important to ensure that carrier recombination occurs at distances well over one diffusion length from surfaces. This can be achieved by making contacts to the p-n junction that are smaller in area than the semiconductor chip or die dimensions. Figure 5.6 shows the effect of a cleaved surface on the luminescence of a
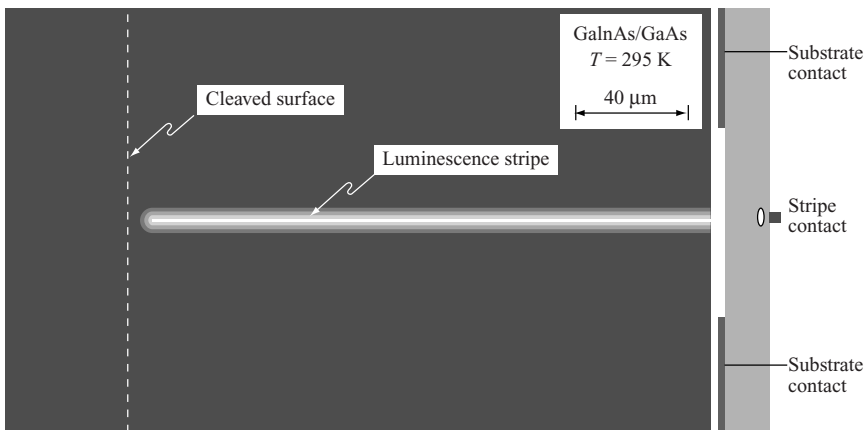
Figure 5.6 *Observed luminescence at the junction of a GaInAs/GaAs LED near a cleaved surface showing that carrier diffusion lengths are in the range of a few microns in GaInAs. Reprinted from E. Fred Schubert, Light-Emitting Diodes, 2e ISBN 978-0-521-86538-8. Copyright (2006) with permission from Cambridge University Press*

LED. Within a few microns of the surface the luminescence falls off rapidly. Surface recombination is relevant to minority carriers and therefore parts of a LED that involve only majority carrier flow are not subject to this issue.

Achieving a balance between electron and hole currents is also important for high-efficiency LEDs. The total current flow in a p-n junction diode can arise from various ratios of holes and electrons crossing the junction; however, a similar number of electrons and holes arriving near the junction per second is needed to maximize recombination efficiency in a LED. This requires that the doping levels on either side of the LED junction must be carefully controlled. In addition a *current blocking layer* may be incorporated to selectively prevent or attenuate the flow of LED current. This can prevent light emission underneath contacts since this light may not be efficiently outcoupled (see Figure 5.4).

5.5   Optical Outcoupling

Due to total internal reflection, only a fraction of the light generated in an LED die will be emitted. Light crossing an interface between two regions of refractive indices n_1 and n_2 must obey Snell's law,

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{n_2}{n_1}$$

In Figure 5.7 the critical angle for light generated in the semiconductor and reaching an air interface occurs when the angle in the air is $\frac{\pi}{2}$. If θ_c is the critical angle in the semiconductor then $n_s\sin\theta_c = 1$ where n_s is the index of refraction of the semiconductor.

Since θ_c is small for practical semiconductors we can use the approximation $\sin\theta_c \cong \theta_c$, and now the critical angle occurs when

$$\theta_c \cong \frac{1}{n_s} \tag{5.2}$$

Using the spherical polar coordinate system in Figure 5.7, an *escape cone* for the light reaching the interface results because all values of ϕ need to be considered. The portion of the surface area of a sphere of radius r that corresponds to radial directions that lie within the escape cone is given by

$$A = \int dA = \int_0^{2\pi} \int_0^{\theta_c} r^2 \sin\theta d\theta d\phi = 2\pi r^2(1 - \cos\theta_c) \tag{5.3}$$
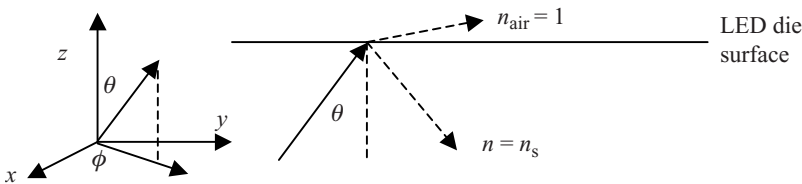
Figure 5.7   *Light generated in the semiconductor will reach the surface and either reflect or be able to exit depending on the angle of incidence. The critical angle is θ_c*

The resulting outcoupling efficiency η_{out} is now the fraction of the sphere through which light may escape:

$$\eta_{\text{out}} = \frac{2\pi r^2(1 - \cos\theta_c)}{4\pi r^2} = \frac{1 - \cos\theta_c}{2} \tag{5.4}$$

Example 5.2

Calculate and compare the optical outcoupling at a GaAs die surface given the following parameters: index of refraction of GaAs $n_s = 3.4$; index of refraction of epoxy $n_e = 1.50$.

(a) in air
(b) in an epoxy encapsulation.

Solution

(a) In air:
    critical angle in air

$$\theta_c = \frac{1}{n_s} = \frac{1}{3.4} = 0.29 \text{ radians or } 16.9°$$

$$\eta_{\text{out}} = \frac{1 - \cos\theta_c}{2} = \frac{1 - \cos(16.9°)}{2} = 0.021 \text{ or } 2.1\%$$

(b) In epoxy resin:
    for critical angle θ_c in epoxy:

$$\frac{\sin 90°}{\sin\theta_c} = \frac{3.4}{1}$$

Solving, $\theta_c = 26.2°$ and

$$\eta_{\text{out}} = \frac{1 - \cos\theta_c}{2} = \frac{1 - \cos(26.2°)}{2} = 0.051 \text{ or } 5.1\%$$

These are poor outcoupling values, and a number of techniques have been developed to improve this, which will be discussed below and in Section 5.11.

It is interesting to examine the dependence on exit angle of the light that does outcouple. If $\theta < \theta_c$ then light will be emitted at an angle Θ, as shown in Figure 5.8.

For light within angle range $d\theta$ the output angle range can be determined using Snell's law:

$$n_s\sin\theta = n_a\sin\Theta \cong n_s\theta$$

or

$$\theta = \frac{n_a\sin\Theta}{n_s}$$
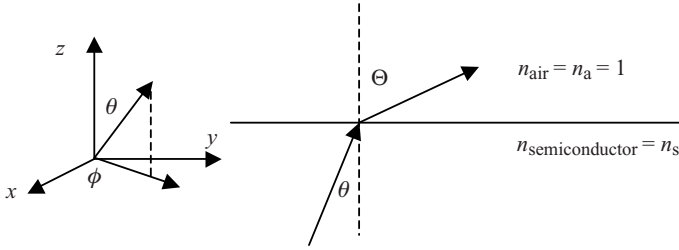
Figure 5.8 *Light that escapes will be diffracted from an incident angle θ to an emitted angle Θ*

Now, differentiating with respect to Θ we obtain

$$\frac{d\theta}{d\Theta} = \frac{n_a}{n_s} \cos \Theta \tag{5.5}$$

If Θ was proportional to θ then the emission profile of an LED chip surface would resemble a point source with equal output of light in all directions; however, the existence of the $\cos \Theta$ term decreases the light being emitted as Θ increases until it reaches zero as Θ approaches 90 degrees. The $\cos \Theta$ dependence of light output results in what is referred to as a *lambertian* source (see Problem 5.19). In a lambertian source luminous intensity falls off with Θ as $\cos \Theta$. Note that since the area of the LED die surface subtended by a viewer at Θ also falls off as $\cos \Theta$, the luminance of the die surface is independent of Θ. See Section 3.9.

The output characteristics of an LED are modified by lenses and reflectors to suit particular applications. Figure 5.3 shows the transparent epoxy polymer lens injected around the LED chip, which helps to increase outcoupling. The polymer has a refractive index of approximately 1.5, which increases the critical angle compared to a die operating in air with refractive index of 1. The outside shape of the epoxy lens can be tailored to achieve a specific beam profile. In many LEDs the amount of light emitted from the sidewalls is greater than the light emitted from the top surface of the LED die. The reflector cup is used to reflect light emitted through the sides of the die towards the lens. A typical emission pattern from an LED mounted as in Figure 5.3 is shown in Figure 5.9.

5.6   GaAs LEDs

A historical view of the development of LEDs is really a history of how direct-gap semiconductors of increasing bandgap were developed. Each successive generation of LED devices required new materials systems, and there have been dramatic improvements in LED performance as new materials became available.

The first type of LED made in the 1960s used GaAs semiconductor material. In Chapter 1 we introduced GaAs as a direct-gap semiconductor with an energy gap of 1.43 eV, which is in the infrared wavelength range. In order to create n-type and p-type GaAs, it is necessary to dope the GaAs with an impurity. Silicon, being a group IV element, is capable of doping GaAs n-type if it substitutes for a Ga atom and p-type if it substitutes for an As atom.

Liquid phase epitaxy (LPE) is a growth technique that has been widely used to grow GaAs and related semiconductors. In LPE, a liquid phase melt containing a solution of Ga
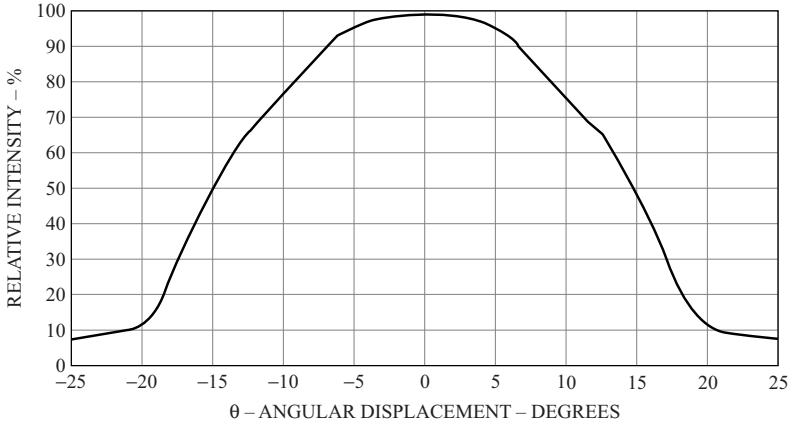
Figure 5.9 *Typical radiation pattern for a LED. This AlGaInP red LED has a radiation pattern with a 30° (±15°) beam divergence determined at half maximum intensity. Reproduced by permission of Avago Technologies from data sheet file AV02-1542EN, http://www1.futureelectronics.com/doc/AVAGO%20TECHNOLOGIES/HLMP-BD16-P0000.pdf, Copyright (2008) with permission from Avago Technologies, USA*

and As atoms is prepared. The Ga concentration in the melt is much higher than the As concentration. In this way the melt consists of Ga solvent and As solute atoms. If such a melt is slowly cooled in contact with a single crystal GaAs substrate a layer of GaAs will nucleate and grow on the substrate as the temperature drop causes the As to become supersaturated in the solution. Note that the grown layer will be single crystal and will follow the crystallographic structure of the substrate.

Si atoms can also be incorporated into the melt of Ga and As. If the epitaxial layer grows at 900°C then the silicon atoms cause the GaAs to be n-type. If the epitaxial layer grows at 850°C the silicon atoms form p-type GaAs. The resulting p-n junction behaves as illustrated in Figure 5.2.

Since GaAs has a small energy gap the photons are infrared with a wavelength of 900 to 980 nm. The resulting device is useful for applications such as remote controls and for short distance optical fibre communications.

5.7 $GaAs_{1-x}P_x$ LEDs

The interest in visible light emission led to the incorporation of phosphorus into LEDs in the 1960s. From Figure 1.33a it is seen that as x increases in $GaAs_{1-x}P_x$ the bandgap increases. Visible emission starts at 750 nm, which corresponds to a bandgap of approximately 1.65 eV and therefore x may be as low as about 0.25 (see Figure 5.10).

The visible emission from these $GaAs_{1-x}P_x$ LEDs is not efficient, as shown in Figure 5.1, for a number of reasons:

(a) As the amount of phosphorus increases the lattice constant increases (see Figure 1.33a). The result is that high-quality epitaxial growth of $GaAs_{1-x}P_x$ on GaAs substrates cannot be achieved unless x is very small. At x values of 0.25 or more, lattice defects and
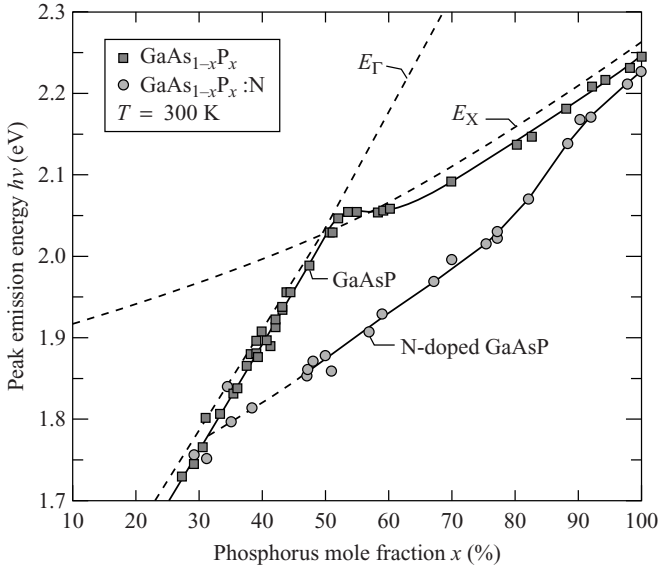
Figure 5.10 *Photon energy plotted as a function of phosphorus mole fraction x in GaAs$_{1-x}$P$_x$ and GaAs$_{1-x}$P$_x$:N. Reprinted with permission from Craford, M. G., et al, Radiative recombination mechanisms in GaAsP diodes with and without nitrogen doping, J. Appl Phys, 43: 10, 4075. Copyright (1972) with permission from American Institute of Physics*

dislocations are present in high concentrations, which lead to non-radiative carrier recombination.

(b) The human eye sensitivity function (luminous efficacy) is weak at long and short wavelengths near the edges of the visible spectrum (see Chapter 3). An efficient red wavelength for human vision is approximately 620 nm, which requires phosphorus concentrations that are high enough to produce a substantial dislocation density.

(c) GaAs$_{1-x}$P$_x$ undergoes a direct–indirect transition near $x = 0.5$, which places a further constraint on bandgap. Even for x values below 0.5, electrons have a chance of being thermally excited into the secondary valley of the conduction band and recombining non-radiatively through an indirect gap transition. This process becomes more serious as the LED junction temperature increases.

(d) A significant fraction of light generated at the junction will be reabsorbed by the GaAs substrate, which has a smaller bandgap than the GaAs$_{1-x}$P$_x$ light emitting material, as well as by GaAs$_{1-x}$P$_x$ material surrounding the junction.

An interesting way to improve the performance of GaAs$_{1-x}$P$_x$ LEDs is through nitrogen doping of the transition region of the p-n junction only. Nitrogen replaces phosphorus atoms in the lattice. Since both nitrogen and phosphorus are in the same column of the periodic table, the nitrogen doping is called an *isoelectronic* defect as it does not act as n- or p-type doping, but instead it provides a recombination centre inside the bandgap of the GaAs$_{1-x}$P$_x$:N material about 100–200 meV below the conduction band minimum. It does this by effectively trapping electrons before recombination occurs. The observed photons therefore have longer wavelengths than from material without nitrogen doping, but because

the nitrogen doping localizes the recombination process x may be increased to well above 0.5, and may be as high as 1.0 in GaP:N.

The explanation as to why the indirect gap transition at $x = 0.5$ does not seriously degrade the performance of GaAs$_{1-x}$P$_x$:N LEDs can be understood from the Heisenberg uncertainty principle. This states that both momentum and position of a particle cannot be measured with certainty at the same time, and the relevant uncertainties obey the relationship $\Delta p \Delta x \geq \frac{\hbar}{2}$. The localization of the electron at the nitrogen impurity substantially reduces Δx compared to a conduction band electron. This results in a high degree of uncertainty in momentum and hence uncertainty in wavevector $\Delta k = \frac{\Delta p}{\hbar}$. The probability of a direct transition is therefore enhanced. In addition nitrogen traps effectively compete with other non-radiative crystal defect traps in the lattice-mismatched material, which increases the probability of radiative transitions. At the same time, the decreased energy of the photons generated compared to the host material leads to less reabsorption of light in the host. The latter point is optimized by N doping that is localized at the junction rather than uniformly doping the entire device with N.

As a result, a range of GaAs$_{1-x}$P$_x$:N LEDs have been in production since the 1970s and continue to be widely used for low-cost LEDs in indicator lamps, numerical readouts and display applications. Offsetting the relatively low efficiency of these LEDs (1–2 lm/W) is the low production cost and high yield. Green emitting GaP:N is particularly popular since its emission near 530 nm is closest in wavelength to the peak of the human eye sensitivity curve at 555 nm for this alloy system.

The availability of high-volume GaP substrate material, which is now routinely used for green-emitting GaP:N devices, provides excellent crystal quality and this further optimizes device performance compared with devices grown on GaAs substrates.

5.8   Double Heterojunction Al$_x$Ga$_{1-x}$As LEDs

Another way to increase the bandgap of GaAs is to partly substitute aluminium for gallium in GaAs. Figure 1.33a shows the dependence of bandgap on x in Al$_x$Ga$_{1-x}$As. Of particular interest is the virtually invariant lattice constant for all values of x, which allows high-quality epitaxial growth on GaAs substrates. The semiconductors are direct gap when $0 < x < 0.45$ and indirect for $0.45 < x < 1.0$. In practice high efficiency requires that the chosen bandgap be several kT smaller than the critical bandgap of approximately 2 eV at the direct/indirect transition point of $x = 0.45$. This is again because thermal energy kT adds energy to electrons causing a portion of them to be excited over into the indirect valley of the conduction band in the Al$_x$Ga$_{1-x}$As system.

In practice very high *external quantum efficiency* may be obtained in this system for active layers with x = 0.35 and emission near 650 nm, which is a deep red colour, although this wavelength is not favourable for the human eye response (see Chapter 3). This limits the efficiency expressed in units of lm W^{-1}, which should not be confused with quantum efficiency. Quantum efficiency refers to the probability of a photon being created per EHP recombination event. External quantum efficiency includes outcoupling efficiency, which further specifies that only the photons that outcouple are counted.

A breakthrough in the performance of these devices in the 1980s was due to the *double heterojunction* device design. Here, the aluminium content is lowered right at the junction
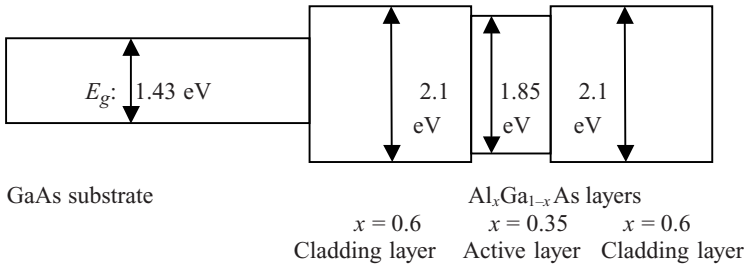
Figure 5.11 *Band gaps of double heterojunction using Al$_x$Ga$_{1-x}$As layers grown epitaxially on a GaAs substrate. Carriers recombine in the active layer of width* W. *The cladding layers are doped such that one layer is n-type and one layer is p-type. Similar structures are used in GaInN LEDs. See Section 5.10*

compared to the remainder of the n- and p-regions. This forms an active layer with reduced bandgap that serves two important purposes:

(a) The active layer captures both electrons and holes in the same spatial region thereby improving the radiative recombination efficiency. Electrons and holes become more likely to encounter one another rather than recombining non-radiatively due to defects, traps and surface states.

(b) Photons generated in the active layer will have a lower energy compared to the surrounding or *cladding* semiconductor material and will therefore not be reabsorbed. Since this material does not need to be direct gap, *x* values well over 0.45 are possible and *x* > 0.6 is common in cladding material.

A diagram of the device structure is shown in Figure 5.11.

The double heterojunction can be approximated as an infinite-walled potential box, as discussed in Section 1.9. The dimensions of the box are the well cross-section area *A* as well as the well width *W*, shown in Figure 5.12. Although a double heterojunction does not have infinitely high walls the infinite-walled model is acceptable provided the electron or hole is confined in energy levels within the wells of heights ΔE_c and ΔE_v respectively, as illustrated.
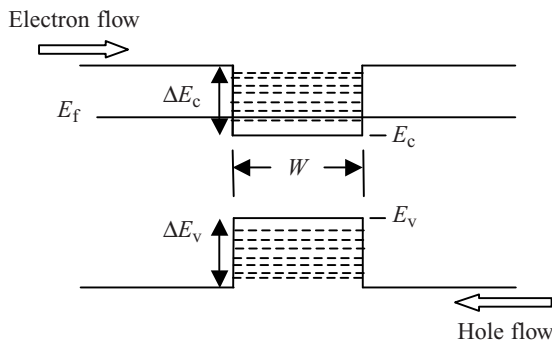
Figure 5.12 *Electron and hole energy levels within the double heterojunction wells of height* ΔE_c *in the conduction band and* ΔE_v *in the valence band*
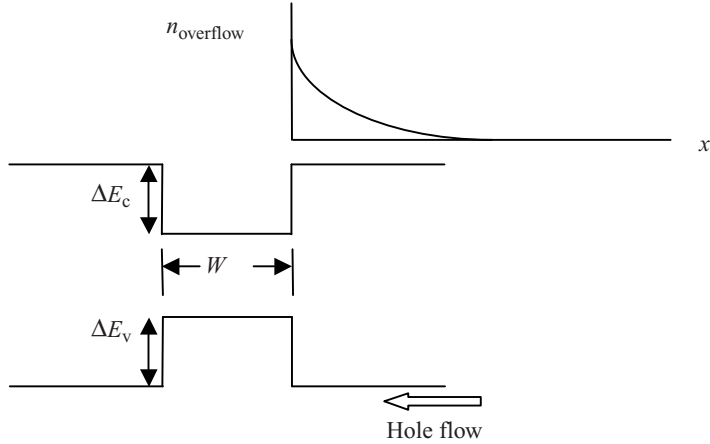
Figure 5.13   *Double heterojunction showing the exponential decay of excess carriers into the cladding layer. Only one side is shown for simplicity*

Firstly we will determine the carrier leakage that can occur in a double heterojunction that is not connected to a current source. In Figure 5.12 the Fermi energy is shown assuming equilibrium conditions. Inside the well it is higher than the bottom of the conduction band. The equilibrium electron concentration is given by Equation 1.32.

If we wish to calculate only the density of *overflow carriers* that are higher in energy than the conduction band of the cladding layer then Equaiton 1.32 must be modified since we wish to consider only carriers higher in energy than $E_c + \Delta E_c$. Hence, the lower limit of the integral in Equation 1.32 is changed from $E_g = E_c$ to become $E_c + \Delta E_c$ and we obtain

$$n_{\text{overflow}} = \int_{E_c + \Delta E_c}^{E_{\text{max}}} D(E - E_g)F(E)\mathrm{d}E$$

The carrier concentration may be calculated using the same method as described in Section 1.9 provided the energy range of interest is well above the Fermi energy. The Boltzmann approximation may be used in place of the Fermi–Dirac distribution. The result is

$$n_{\text{overflow}} = N_c \exp\left(\frac{-(E_c + \Delta E_c - E_f)}{kT}\right) \tag{5.6}$$

which is the same as Equation 1.34 but with E_c replaced by $E_c + \Delta E_c$. The overflow carriers see no energy barrier and can directly diffuse into the cladding layers. This will lead to an exponential decay in concentration as a function of distance into the cladding layers, as shown in Figure 5.13.

Using the diffusion equation from Section 1.18, we have

$$n(x) = n_{\text{overflow}} \exp\left(\frac{-x}{\sqrt{L_n}}\right)$$

and hence at the edge of the well where diffusion current is a maximum we can evaluate the leakage current as

$$J_n(x)_{\text{diffusion}}\,|_{x=0} = q\,D_n\frac{dn(x)}{dx}\,|_{x=0} = -q\frac{D_n}{L_n}n_{\text{overflow}} \tag{5.7}$$

Since both n_{overflow} as well as $\frac{D_n}{L_n}$ vary exponentially with temperature it is clear that the current leakage is also an exponential function of temperature.

Example 5.3

Calculate the room temperature leakage current density in an $Al_xGa_{1-x}As$ double heterostructure without forward bias given the following parameters:

Inside the well $E_f - E_c = 50\,\text{meV}$
Barrier height $\Delta E_c = 200\,\text{meV}$
Effective density of states $N_c = 5 \times 10^{17}\,\text{cm}^{-3}$
Carrier lifetime $\tau = 5 \times 10^{-9}\,\text{s}$
Carrier mobility $\mu_n = 1500\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1}$

Solution

Using Equation 5.6:

$$n_{\text{overflow}} = N_c \exp(-(E_c + \Delta E_c - E_f)/kT)$$
$$= 5 \times 10^{17}\,\text{cm}^{-3}\exp\left(\frac{-150\,\text{meV}}{26 \times 10^{-3}\,\text{eV}}\right) = 1.56 \times 10^{15}\,\text{cm}^{-3}$$

and

$$\frac{D_n}{L_n} = \frac{D_n}{\sqrt{D_n\tau_n}} = \sqrt{\frac{D_n}{\tau_n}}$$

Using the Einstein relation,

$$D_n = \frac{kT}{q}\mu_n = 26 \times 10^{-3}\,\text{V} \times 1500\,\text{cm}^2\,\text{V}^{-1}\,\text{s}^{-1} = 39\,\text{cm}^2\,\text{s}^{-1}$$

Now from Equation 5.7:

$$J_n(x)_{\text{diffusion}}|_{x=0} = q\frac{D_n}{L_n}n_{\text{overflow}} = q\sqrt{\frac{D_n}{\tau_n}}n_{\text{overflow}}$$

$$= 1.6 \times 10^{-19}\,\text{C} \times \sqrt{\frac{39\,\text{cm}^2\,\text{s}^{-1}}{5 \times 10^{-9}\,\text{s}^{-1}}} \times 1.56 \times 10^{15}\,\text{cm}^{-3} = 22.0\,\text{A cm}^{-2}$$

To put this result in perspective it should be compared to a typical LED current density of approximately 100–$1000\,\text{A cm}^{-2}$. This indicates that carrier leakage can be a significant factor especially at higher operating temperatures, and ΔE_c must be large enough to prevent excessive carrier overflow.

In an operating double heterojunction LED we also need to consider the flow of current and the *rate* at which carriers arrive at the conduction and valence band wells.

If an electron current density J flows in steady state into the conduction band well as shown in Figure 5.12, the current adds to the electron population of the well. In one second the number of electrons added to the well will be $\frac{JA}{q}$. Since the volume of the well is AW the number of electrons added per second per unit volume in the well is $\frac{J}{Wq}$. This can be expressed in differential form as

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \frac{J}{Wq} - R \tag{5.8}$$

where R is the recombination rate of the carriers in the well.

As discussed in Section 1.15, the recombination rate $R \propto np$, and we introduced low-level injection conditions. In the case of low-level injection we either use a fixed recombination time τ_n, which assumes that minority electrons recombine with majority holes having a constant hole carrier concentration, or a fixed recombination time τ_p which assumes that minority holes recombine with majority electrons having a constant electron carrier concentration

For a forward-biased LED we can no longer assume low-level injection conditions because both hole *and* electron concentrations will be at high levels, neither of which are fixed in the active region when typical forward currents of $100\,\mathrm{A\,cm^{-2}}$ or more are flowing. The recombination rate $R \propto np$ is valid but the recombination time is not fixed in high-level injection; it decreases as the carrier concentration increases.

To model this we introduce a *recombination coefficient B* such that

$$R = Bnp \tag{5.9}$$

Values of B for direct-gap III-V semiconductors are approximately $10^{-10}\,\mathrm{cm^3\,s^{-1}}$, whereas in indirect gap semiconductors such as GaP, Si and Ge the values of B are three to four orders of magnitude smaller than this since the participation of a photon is generally required for momentum conservation.

Substituting Equation 5.9 into Equation 5.8, and assuming that the supply of holes and electrons is balanced ($n = p$), we obtain

$$n = \sqrt{\frac{J}{qBW}} \tag{5.10}$$

The number of available states in an infinite-walled well was derived in Chapter 1 and can be used to estimate how many electron states the well can accommodate. If we regard the bottom of the upper well as zero in energy and assume complete filling of all available states up to the top of the well then we can identify ΔE_c as the highest energy level E in the well. Evaluating Equation 1.30 for $E = \Delta E_c$ and using effective mass m_e^* we obtain

$$n = n(E) = \frac{\pi}{3}\left(\frac{2m_e^*\Delta E_c}{\hbar^2\pi^2}\right)^{\frac{3}{2}} \tag{5.11}$$

Equating 5.10 and 5.11, and solving for J, which now represents the maximum current density that can be accommodated before the well overflows, we obtain

$$J_{\max} = qBW\left(\frac{\pi}{3}\right)^2\left(\frac{2m_e^*\Delta E_c}{\hbar^2\pi^2}\right)^3 \tag{5.12a}$$

and using a similar argument for the lower well we obtain

$$J_{\max} = qBW \left(\frac{\pi}{3}\right)^2 \left(\frac{2m_{\mathrm{h}}^* \Delta E_{\mathrm{v}}}{\hbar^2 \pi^2}\right)^3 \tag{5.12b}$$

The well that overflows first effectively limits the current density.

This shows that the maximum current flow that can be accommodated in the well is proportional to the well width and the cube of the well height. For high-current-density LEDs a wide well is required. Typically several thousand amps per square centimetre can be accommodated in a double heterostructure LED.

Example 5.4

Calculate the current density level at 300 K at which the electron well overflows in an $Al_x Ga_{1-x}As$ double heterostructure with the following parameters:

Barrier height $\Delta E_{\mathrm{c}} = 200\,\mathrm{meV}$
Well width $W = 200\,\mathrm{A}$
Effective density of states $N_{\mathrm{c}} = 5 \times 10^{17}\,\mathrm{cm}^{-3}$
Recombination coefficient $B = 2 \times 10^{-10}\,\mathrm{cm}^3\,\mathrm{s}^{-1}$

Solution

From Equation 1.34b:

$$N_{\mathrm{c}} = 2 \left(\frac{2\pi m_{\mathrm{e}}^* kT}{\hbar^2}\right)^{\frac{3}{2}}$$

Solving for m_{e}^* we obtain

$$m_{\mathrm{e}}^* = \frac{\hbar^2}{2\pi kT} \left(\frac{N_{\mathrm{c}}}{2}\right)^{\frac{2}{3}} = \frac{(6.62 \times 10^{-34}\,\mathrm{Js})^2}{2\pi (0.026 \times 1.6 \times 10^{-19}\,\mathrm{J})} \left(\frac{5 \times 10^{17}\,\mathrm{cm}^{-3} \times 10^6\,\mathrm{cm}^3\,\mathrm{m}^{-3}}{2}\right)^{\frac{2}{3}}$$
$$= 6.63 \times 10^{-32}\,\mathrm{kg}$$

Now,

$$J_{\max} = qBW \left(\frac{\pi}{3}\right)^2 \left(\frac{2m_{\mathrm{e}}^* \Delta E_{\mathrm{c}}}{\hbar^2 \pi^2}\right)^3$$
$$= 1.6 \times 10^{-19}\,\mathrm{C} \times 2 \times 10^{-10}\,\mathrm{cm}^3\,\mathrm{s}^{-1} \times 10^{-6}\,\mathrm{m}^3\,\mathrm{cm}^{-3} \times 200 \times 10^{-10}\,\mathrm{m}$$
$$\times \left(\frac{\pi}{3}\right)^2 \left(\frac{2 \times 6.63 \times 10^{-32}\,\mathrm{kg} \times 0.2\,\mathrm{eV} \times 1.6 \times 10^{-19}\,\mathrm{J\,eV}^{-1}}{(1.05 \times 10^{-34}\,\mathrm{Js})^2 \times \pi^2}\right)^3$$
$$= 4162\,\mathrm{A\,cm}^{-2}$$

This shows that this double heterojunction functions well at high current densities.

Since $J_{\max}$ decreases as W decreases, a very narrow double heterojunction well of width W under about 100 Å, which is referred to as a *quantum well,* will overflow at considerably

lower current densities. The direct evidence for this overflow is an observed saturation in quantum well LED brightness at a current density that is lower than that of the same LED but having a wider well. In a quantum well LED the number of available states in the well is small; however, more than one quantum well may be used in an LED. There are advantages to the use of quantum wells and this is discussed further in Section 5.10.

Still higher efficiencies are possible if the absorbing GaAs substrate can be removed. Techniques have been developed whereby after the epitaxial layers are grown, a wide-bandgap foreign substrate is bonded to the top surface of the device whereupon the original GaAs substrate is chemically or mechanically removed. A technique known as *wafer bonding* is used to bond the grown LED structures to a transparent substrate such as GaP, which provides a stable and transparent support for the epitaxial layers before the GaAs substrate is removed. The wafer bonding technique uses heat and pressure to form the bond but does not require that the new substrate be lattice matched with the active layers. Efficiencies close to $10 \, \text{lm} \, \text{W}^{-1}$ are achieved in practice in combination with the double heterostructure.

One challenge in high-aluminium-content $\text{Al}_x\text{Ga}_{1-x}\text{As}$ material is the tendency for aluminium to react with oxygen and moisture in the surrounding atmosphere. For this reason the lifetimes of these devices can be limited unless they are properly protected.

5.9   AlGaInP LEDs

The leading LED material system for high-brightness red, orange and amber/yellow LEDs is the quaternary system AlGaInP. It was developed in the late 1980s and early 1990s, and permits an additional degree of freedom in controlling bandgap and lattice constant compared to ternary alloys. By varying the composition both the bandgap and the lattice constant may be varied independently. The lattice constant is maintained to match that of GaAs while the bandgap may be direct gap up to 2.33 eV. This has yielded orange LEDs with $100 \, \text{lm} \, \text{W}^{-1}$ efficiency by incorporating a double heterostructure. A more detailed picture of the composition range is given in Figure 5.14.

The preferred growth method for these quaternary materials is *metal-organic vapour phase epitaxy* (MOVPE), also referred to as *metal-organic chemical vapour deposition* (MOCVD). Unlike the use of the liquid melt used in LPE growth, MOVPE growth is a gas-phase growth technique in which molecules containing the required semiconductor atoms are allowed to pass over a heated substrate in a gas stream usually combined with an inert gas such as argon. The molecules contain an organic group composed of elements hydrogen and carbon in combination with one of the desired Al, Ga, In or P atoms (or the required dopant atom). The gas stream is composed of a well-controlled mixture of the desired molecules. By carefully establishing the ratios of the respective molecules in the gas stream the desired semiconductor material is grown on the substrate because the substrate temperature is high enough to crack the molecules allowing the semiconductor atoms to deposit on the substrate while the remaining organic species flow away with the gas stream.

Advantages of MOCVD growth include the ability to grow ternary and quaternary semiconductors covering a wide composition range while simultaneously achieving a high degree of control of composition. During growth, composition changes may be made by changing the flow rates of specific molecular species.
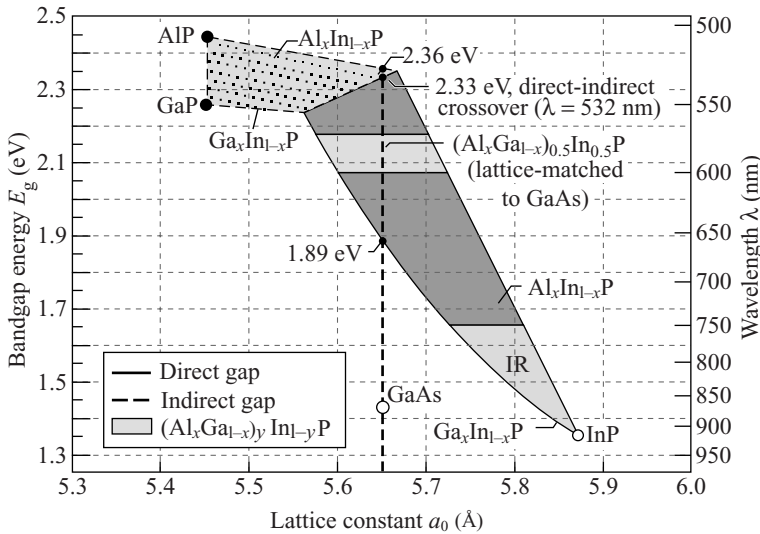
Figure 5.14 *(Al_xGa_{1−x})_yIn_{1−y}P bandgap versus lattice constant graph showing the composition ranges in this quaternary system. By adjusting the two available parameters, x and y, a field of compositions is possible represented by the shaded areas. A range of energy gaps from 1.89 eV to 2.33 eV is available in the direct bandgap region while matching the GaAs lattice constant. Reproduced with permission from Elsevier from OMVPE Growth of AlGaInP for High-Efficiency Visible Light-Emitting Diodes, Semiconductors and Semimetals, Volume 48, C.H. Chen, S.A. Stockman, M.J. Peanasky, C.P. Kuo Copyright (1997) Elsevier Ltd*

Although the highest achievable bandgap of 2.33 eV in the AlGaInP system corresponds to green emission at 532 nm, efficient green emission is not available from this system because of the need to reduce bandgap several kT below the direct–indirect transition point. In addition, the bandgaps of this quaternary system are sensitive to ordering of the atoms. Figure 5.14 is for unordered material in which the Al, Ga and In atoms are randomly positioned in the lattice. Unordered material is not always obtainable since there is a thermodynamic drive for atoms to order during growth to minimize energy. Atomic ordering can lower the band energy of the crystal and the degree of ordering is dependent on composition. It is experimentally found that when ordering occurs the effective bandgap drops by as much as 190 meV. For these reasons yellow is the shortest practical emission wavelength for this system.

The efficiency peaks in the orange wavelength range since this represents the best compromise between the human eye response (luminous efficacy), which falls off as wavelength increases above 555 nm, and the recombination efficiency, which falls off as wavelength decreases and gets closer to the direct–indirect transition. The ability of the quaternary AlGaInP system to reach shorter wavelengths with high efficiency results in an order of magnitude improvement in luminous efficiency compared to the $Al_xGa_{1−x}As$ system.

As with AlGaAs LEDs the double heterostructure is used to improve the probability of radiative electron-hole recombination but in addition larger values of ΔE_c and ΔE_v in the heterostructure can be achieved due to the bigger bandgap range available, which further improves carrier confinement even at higher operating temperatures.

For the highest efficiencies the removal of the GaAs growth substrate is required since GaAs absorbs a portion of the emitted light. Wafer bonding is employed as well as further outcoupling improvements using tilted facets, as discussed in Section 5.11.

5.10 $Ga_{1-x}In_xN$ LEDs

Since the early 1990s the ternary nitride materials system $Ga_{1-x}In_xN$ capable of yielding high-efficiency LEDs covering the green and blue parts of the visible spectrum has been developed and widely commercialized using nitride semiconductors.

Prior to this breakthrough using nitrides, II-VI semiconductors were thought to be ideal candidates for green and blue LEDs due to their direct bandgaps. These materials, including direct-gap semiconductors ZnO, CdS and CdSe, did not demonstrate sufficiently stable performance partly due to the difficulty in incorporating p-type dopants. For the substitution of a group II metal (Zn or Cd) by a p-type dopant the choices consist of group I metals such as sodium or other singly ionized transition metals such as copper. Unfortunately, singly ionized atoms are only weakly bonded in II-VI crystals and they are therefore mobile and hence unstable.

The development of GaN-based LEDs was hampered by a number of challenges prior to the 1990s, and work on them was limited and confined to small research groups. The challenges with the growth and successful doping of these materials included the following:

(a) GaN as shown in Figure 1.30b is not lattice matched with readily available single crystal substrates.
(b) The growth conditions for crystalline GaN by MOCVD require high-temperature growth of 900°C or more whereas other non-nitride III-V LEDs are grown at temperatures between 400 and 600°C. This further limits the range of suitable substrates.
(c) Doping to achieve p-type material is required and had not been successfully demonstrated and understood.

GaN is a direct-gap semiconductor closely related to the family of other III-V semiconductors used for LEDs, although it has the wurtzite crystal structure (see Sections 1.10 and 1.11). GaInN has now proved to be the most viable approach to large, direct bandgap materials for LEDs in spite of the challenges. These challenges have been overcome, allowing dramatic new results to be obtained, and $Ga_{1-x}In_xN$ alloy semiconductors are now established as being of primary importance to LED technology.

Using an MOCVD growth method, smooth, high-quality films of GaN may be grown on substrates that are not lattice matched with GaN. The most important substrates are sapphire and SiC, which are mismatched by approximately 12% and 3% respectively. The achievement of high-performance GaN material in spite of the poor lattice match has been the subject of much investigation. It is now understood to be possible due to the formation of dislocations in the GaN that form near the surface of the substrates but that reduce in density as the GaN film grows thicker. The dislocation densities achieved of 10^7 to 10^9 cm^{-2} are still high compared to the dislocation densities of below 10^6 cm^{-2} achievable in non-nitride III-V materials; however, they are not as detrimental to LED performance as was expected. A comparison is shown in Figure 5.15, which plots the observed efficiency drop as a function of dislocation density for a variety of semiconductors. The nitrides show
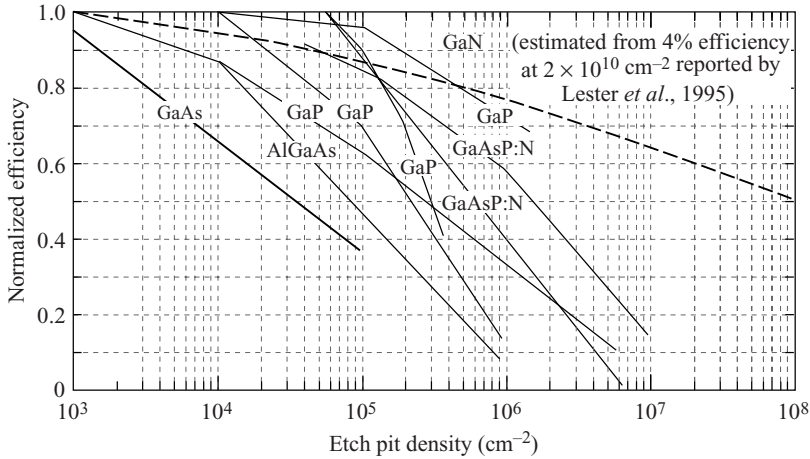
Figure 5.15 *Radiative efficiency as a function of dislocation (etch pit) density for a variety of III-V semiconductors. Dislocation density is determined by etching the crystal surface and then counting the number of resulting etch pits per unit area. Etch pits form at the dislocations. Reprinted from E. Fred Schubert, Light-Emitting Diodes, 2e ISBN 978-0-521-86538-8. Copyright (2006) with permission from E. Fred Schubert*

a significantly higher tolerance to dislocation density. Figure 5.16 illustrates the density of dislocations on a sapphire substrate and how this dislocation density drops with further growth. By growing the highly faulted layer at lower temperatures the dislocation density in the subsequently grown material is greatly reduced. Even this reduced dislocation density would be too high for all but the nitride materials.
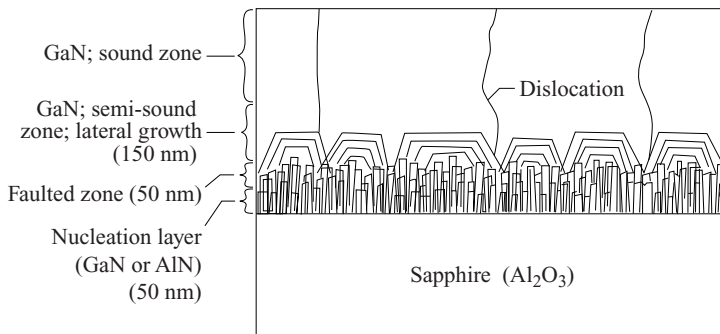
Figure 5.16 *Dislocations in GaN epitaxial layer grown on sapphire. In addition to these factors, the 12% lattice mismatch of GaN with respect to sapphire is effectively much less apparent since it turns out that a rotation about the c-axis of GaN relative to the sapphire substrate allows a far better lattice match of the GaN system relative to the sapphire in the plane normal to the c-axis. See Problem 5.18. Both sapphire and SiC are very stable substrate materials that may be heated to over 1000°C during GaN growth, and both substrates are used in the high-volume production of GaInN LEDs. Reprinted with kind permission from The Blue Laser Diode, S. Nakamura, S. Pearton, G. Fasol, page 14 fig 2.2, Copyright (2000) Springer Science & Business Media*

The issue of doping also had to be resolved. GaN is readily doped n-type using group IV atoms Si or Ge to substitute for Ga. p-Type doping is achieved with Mg, although in other III-V semiconductors such as GaAs, both Zn and Mg are used as p-type dopants. The Mg^{2+} ion is very stable and it can substitute for Ga in spite of the fact that Mg^{2+} is usually incorporated into crystals with six nearest neighbours (octahedral symmetry) rather than the four nearest neighbours (tetrahedral symmetry) found in GaN. The effective p-type doping of GaN with Mg was not initially possible. Eventually a high-temperature anneal after dopant incorporation at temperatures near $900°C$ was discovered to activate the dopant and allow it to act as a stable shallow acceptor. The ability to dope GaN p-type became the final key to the success of the GaInN system.

GaN has a bandgap that is higher than needed for visible light emission, and indium is added to form the alloy semiconductor $Ga_{1-x}In_xN$. The incorporation of In decreases the bandgap and it may be added in the appropriate proportion for blue or green light emission using the MOCVD growth process. In addition In plays a role in reducing the requirement for crystal perfection. It is believed that In incorporation in GaN effectively traps electrons locally and assists in the recombination process, competing effectively with traps due to dislocations.

The double heterojunction structures that were used for red and yellow LEDs are directly applicable to GaInN LEDs. Figure 5.17 shows the emission spectra of blue and green heterojunction GaInN LEDs as well as a red LED. The red LED is made from AlGaInP. Red-emitting GaInN LEDs can be produced by increasing the In content; however, they are not as efficient as red AlGaInP LEDs. The colours red, green and blue are needed for large, full-colour LED displays, which are a major application of LEDs, and the success of the GaInN system in conjunction with earlier AlGaInP devices effectively revolutionized the use of LEDs for many applications, including full-colour outdoor video displays and trichromatic (red, green, blue) LCD backlighting.
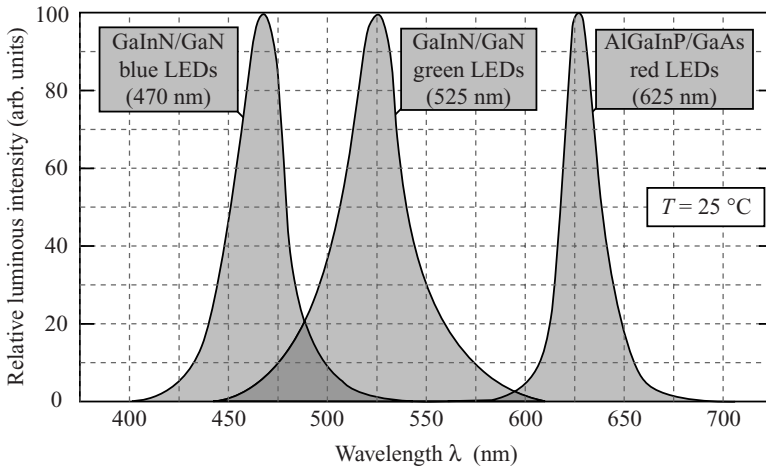
Figure 5.17   *Emission spectra of blue, green and red LEDs having the highest available effi-ciencies. Reprinted with permission of Toyoda Gosei*
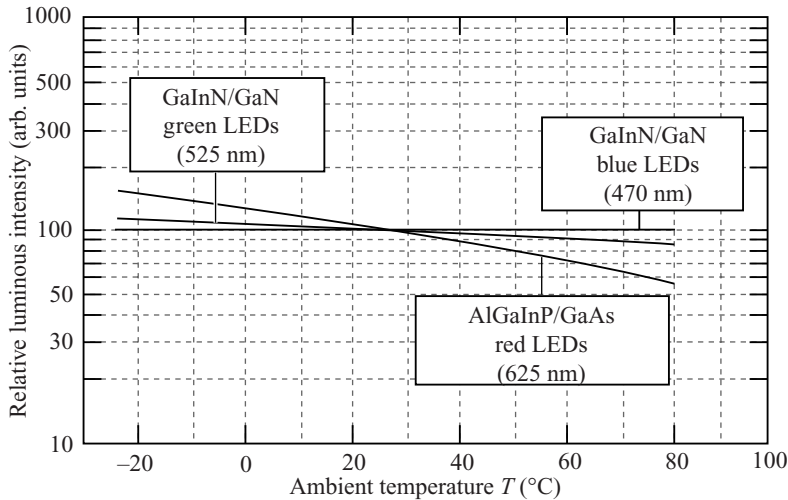
Figure 5.18   *Output intensity versus ambient temperature for GaInN and AlGaInP LEDs. Note the decreased thermal quenching in GaInN. Reprinted with permission of Toyoda Gosei*

The temperature behaviour of LED emission intensity depends on the semiconductor material as shown in Figure 5.18. AlGaInP devices show stronger thermal quenching because of electrons that are thermally excited into the secondary valley indirect gap recombination process, whereas in InGaN alloys there is no indirect gap. Also the double heterostructure used in the GaInN devices has a larger energy barrier between the active layer and the cladding layers, which maintains carrier confinement at higher operating temperatures than in AlGaInP.

The degree of ideality in the performance of blue- and especially green-emitting GaInN LEDs falls short of that achieved in AlGaInP devices, as shown in Figure 5.19. The lower steepness of the intensity versus current characteristic in the nitrides is indicative of a higher series resistance. One serious contribution is the higher resistance of the p-type GaN. Not all the Mg dopants are ionized since the acceptor level is approximately 200 meV away from the valence band, which limits the hole concentration. Also in green LEDs the additional In content leads to added heat as carriers relax before recombination takes place. This prevents the forward voltage from being as low as predicted for green emission.

The dependence of bandgap energy on lattice constant for the alloy semiconductor systems $Al_xGa_{1-x}N$, $Ga_{1-x}In_xN$ and $Al_xIn_{1-x}N$ are shown in Figure 5.20. Note that the lines joining the binary compounds are actually bowed rather that straight as shown in Figure 1.33b.

GaN has a hexagonal crystal structure and is usually grown in the *c*-axis direction such that close-packed planes alternating between Ga and N atoms lie perpendicular to the growth direction (see Figure 1.15c). Since these atomic planes contain alternating charge densities, with positively charged Ga planes and negatively charged N planes, there is a natural polarization due to charge distribution along the *c*-axis of the crystal. If these opposing charge densities are equal and opposite there will be no *average* polarization
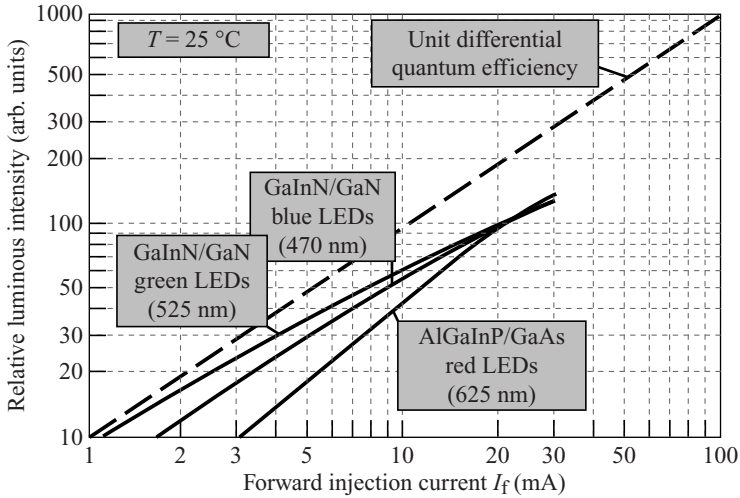
Figure 5.19   *Forward intensity versus current characteristics for GaInN and AlGaInP LEDs. Reprinted with permission of Toyoda Gosei*

along the *c*-axis; however, in practice there is a significant average polarization, which depends strongly on compressive strain in the crystal. Consider the active layer in the double heterostructure of Figure 5.21a. The In added to the active layer induces lateral compressive strain in the active layer, which significantly contributes to a net polarization field along the *c*-axis. This means that the energy band diagram must be modified as shown in 5.21b. When an electron and a hole are injected into the active layer they are naturally pulled to opposite sides of the layer and a reduced radiation efficiency results. This difficulty can be overcome by making the active layer thinner, and layer thicknesses in the range of 20 Å are suitable.

These devices are quantum well LEDs and suffer from low maximum current densities due to the narrow well width (see Equation 5.12). This difficulty can be addressed by increasing the number of quantum wells to form a *multiple quantum well* LED. Devices with six or eight quantum wells are common, and these structures are now in high-volume production. The MOCVD growth technique is capable of producing accurate and reproducible multiple quantum well structures.

Another way to address the issue of polarization and the unwanted separation of electrons and holes is to grow GaN layers in a different crystallographic orientation. Figure 5.22 shows other growth planes that may be used. These alternative planes are either non-polar or semi-polar, and LEDs grown with these orientations have demonstrated beneficial carrier recombination efficiencies and better quantum efficiencies at high current densities. Mass production of LEDs using these non-polar or semi-polar structures is not yet in place. Substrates and new growth techniques suited to non-*c*-axis growth are under development.

The desire for white emission from LEDs has motivated the development of LEDs that convert blue light into white light. Since blue-emitting $In_xGa_{1-x}N$ LEDs are highly
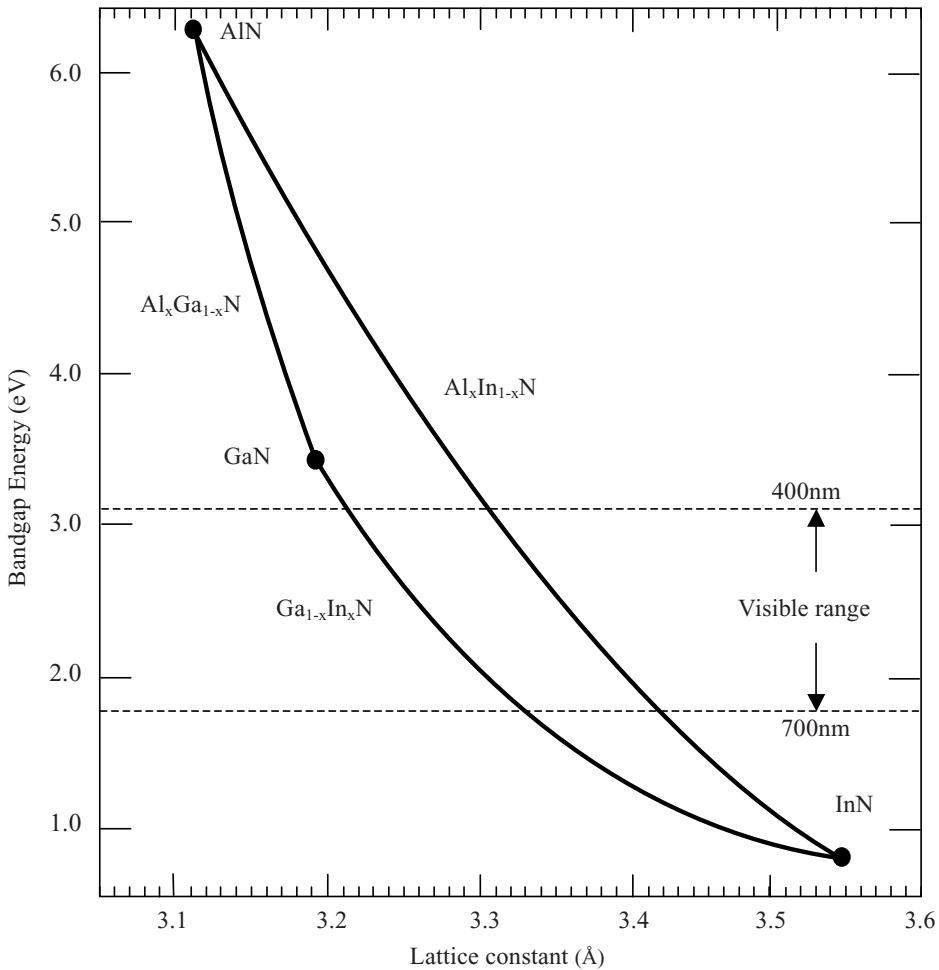
Figure 5.20 *The nitride alloy semiconductor systems $Al_xGa_{1-x}N$, $Ga_{1-x}In_xN$ and $Al_xIn_{1-x}N$ plotted to show energy gap as a function of lattice constant. Reproduced from Schubert EF.* Light Emitting Diodes*, 2nd edn. Cambridge University Press, 2006, p. 223, with permission*

efficient, high-efficiency white LEDs may be achieved by *down-conversion* in which a phosphor material situated in close proximity to the blue-emitting GaInN die absorbs blue LED emission and then re-emits light by phosphorescence at a longer wavelength. The material most widely used to achieve this is YAG:Ce which is abbreviated notation for cerium-doped $Y_3Al_5O_{12}$ phosphor material, which has the garnet crystal structure. This phosphor has strong absorption in a narrow band at 460 nm and broadband yellow emission centred at 550 nm. A fine powder of this phosphor can be included in the epoxy-encapsulated LED package such that a portion of the blue light emitted from the LED die is converted to yellow emission and the remaining blue emission mixes with the yellow to create white light. Figure 5.23 shows the emission spectrum of an LED of this type. The CIE diagram

GaN   In$_x$Ga$_{1-x}$N   GaN
x = 0   x = 0.2   x = 0
Cladding layer   Active layer   Cladding layer
(a)

GaN   In$_x$Ga$_{1-x}$N   GaN
x = 0   x = 0.2   x = 0
Cladding layer   Active layer   Cladding layer
(b)
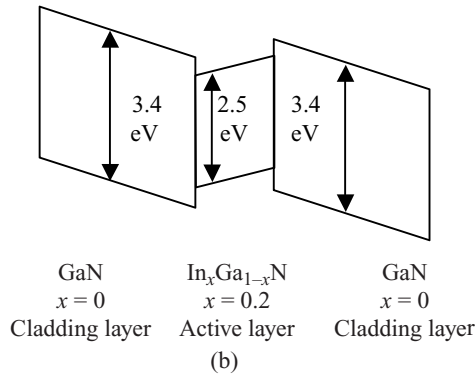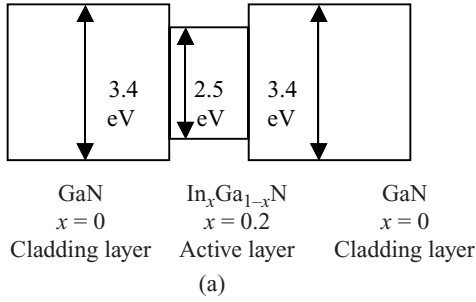
Figure 5.21   *(a) Band diagram of a double heterostructure using In$_x$Ga$_{1-x}$N active layer grown epitaxially. Carriers recombine in the active layer . The cladding layers are doped such that one layer is n-type and one layer is p-type. (b) Band diagram including the effect of polarization*

(Figure 3.13) can be used to understand the achievement of white light emission based on an appropriate combination of blue and yellow light.

The optimization of white-emitting LEDs is an area of ongoing effort. The ideal white light source for many lighting applications is the solar spectrum shown in Figure 4.6, and a deficiency in wavelengths around 500 nm is clearly evident in the spectrum of Figure 5.23. There are a large number of phosphor materials that have specific emission spectra. Also the option exists of using two or more LED die emitting at two or more distinct wavelengths to create a white spectrum.

The requirements for the achievement of a successful white light source include:

(a) Ideality of white light spectrum. This is measured by a *colour rendering index* (CRI), which quantifies the ability of a white light source to satisfactorily illuminate samples having a wide range of colours.
(b) Uniformity of the colour as a function of emission angle.
(c) Stability of the emitted colour. If two or more LEDs are used there could be *differential aging* in which the relative intensities of the two or more die change.
(d) Cost of materials and assembly.
(e) Efficiency of the resulting white light source.

Non-Polar Planes　　　　　　　**Semi-Polar Planes**
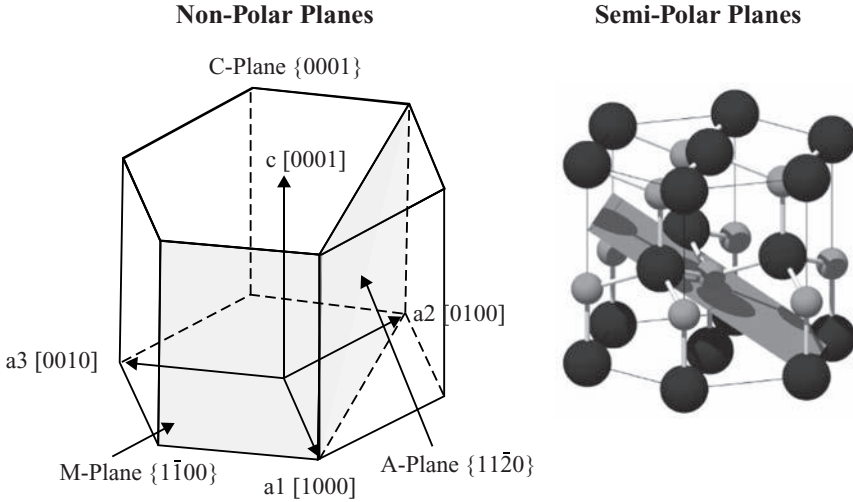
Figure 5.22 *GaN planes are generally (0001) planes that are formed as a result of c-axis growth; however, alternative growth directions and planes may also be achieved. This reduces or eliminates polarization in the growth directions. M-planes (1$\bar{1}$00) or A-planes (1$\bar{1}$20) as shown are non-polar and semi-polar planes are also available. Growth of high quality GaN in directions resulting in non-polar and semi-polar quantum well structures is an area of current LED development. Reprinted from Speck, J. S., New Faces of GaN: Growth, Doping and Devices from INSIGHTS 2006, http://engineering.ucsb.edu/insights2006/watch.php?video=speck. Copyright (2006) with permission from James Speck*
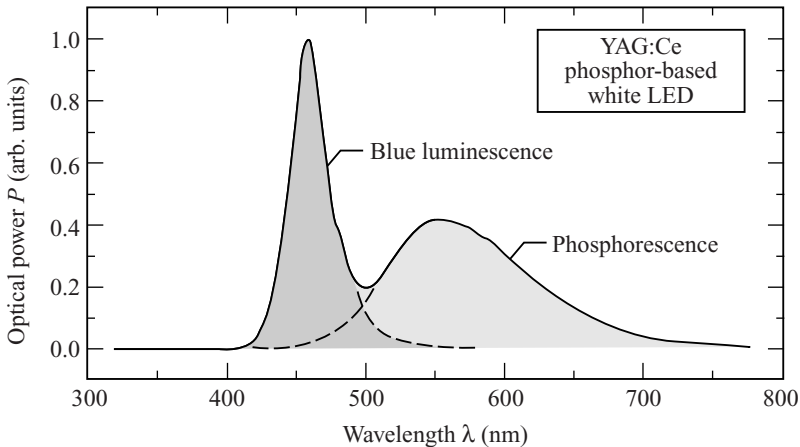
Figure 5.23 *Emission spectrum of white-emitting LED. Blue light from the LED die is down-converted using YAG:Ce to produce a broadband yellow emission. When combined with the blue LED emission white light results. Reprinted with permission from Nichia Corp*

5.11   LED Structures for Enhanced Outcoupling and Power Output

The low outcoupling inherent in LED structures may be improved beyond what is discussed in Section 5.5. As shown in Figure 5.24 there are six possible escape cones for the light, one for each facet of the die. It is usual that the top and bottom surfaces of the die have contacts that either partially block or fully block light trying to escape, which reduces the overall outcoupling.

If the side cuts of the die are tilted then outcoupling can be improved. This is illustrated in Figure 5.25. Light that internally reflects from the top face may reach the tilted facets at an angle smaller than the critical angle and outcoupling can take place. The best tilt angle is found to be $\theta = 35°$ in AlGaInP LED material, and the resulting quantum efficiency of AlGaInP LEDs made using tilted facets exceeds $100\,\mathrm{lm\,W^{-1}}$ for an orange-emitting device with peak wavelength of 610 nm. An improvement in outcoupling by a factor of 1.4 has been observed compared to vertical sidewall LEDs. Further outcoupling improvements can be realized if spherical or cylindrical LED die are formed. In these structures light from the active region has a high likelihood of arriving almost normal to the die surface; however, these structures are costly to fabricate and require deep etching steps.

Surface texturing has also been shown to improve optical outcoupling. Just as texturing is routinely used to decrease reflection losses in solar cells, as discussed in Chapter 4, it can also be applied to LEDs. This is illustrated in Figure 5.26.

The *distributed Bragg reflector* (DBR) is a means of reflecting generated light away from the LED substrate and preventing potential absorption in the substrate or back contact. It is shown in Figure 5.4. A stack of alternating layers of semiconductors with two different values of refractive index is designed to constructively add the reflection of light at each interface together to provide strong reflection at the wavelength of LED emission. Semiconductor refractive index varies with composition and therefore ternary and quaternary III-V semiconductors lend themselves to form efficient DBR structures without compromising crystal quality (see Problem 5.17). The DBR adds cost to die fabrication and is not widely used for low-cost LEDs.

A small LED has its die cut in the range of 200–400 μm in both length and width, and has a junction area of under $10^{-4}\,\mathrm{cm^2}$, a maximum steady current of 30–50 mA or current density under about $100\,\mathrm{A\,cm^{-2}}$. These LEDs do not require a heatsink and are ideal for a
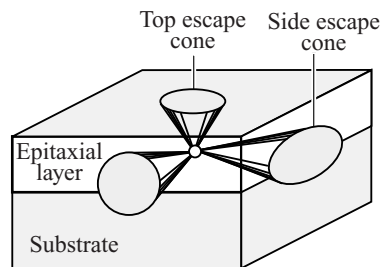
Figure 5.24   *Three of the six possible escape cones for light emission from a LED die with vertical sidewalls. Reprinted from E. Fred Schubert, Light-Emitting Diodes, 2e ISBN 978-0-521-86538-8. Copyright (2006) with permission from Cambridge University Press*
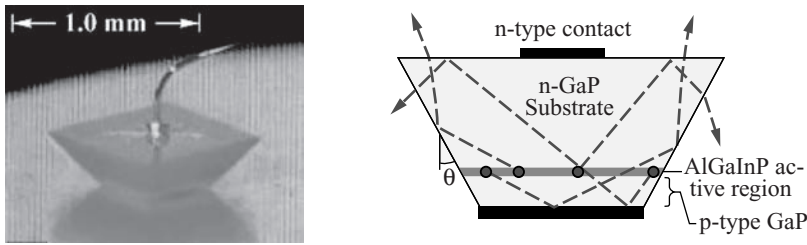
Figure 5.25 *Light can outcouple more efficiently if the sidewalls of the die are tilted as shown. The tilted walls can be applied to a variety of substrates. Reproduced from M.R. Krames et al., High-power truncated-inverted-pyramid (AlxGa1-x)0.5In0.5P/GaP light-emitting diodes exhibiting > 50% external quantum efficiency, Applied Physics Letters Vol. 75, No. 16. Copyright (2000) with permission from American Institute of Physics*

wide range of applications including pixels on outdoor video signage, LCD backlighting and low-to-medium-level illumination applications.

High-brightness high-power LEDs represent a growing sector of LED technology. These LEDs have junction areas that are 10, 100 or even 1000 times the area of small LEDs and operate at up to about $1000\,A\,cm^{-2}$. They generally do require heatsinking as well as specialized packaging and light outcoupling schemes.

Figure 5.27 shows the packaging structure of a high-power LED that is capable of generating a luminance of over 300 lm from a single die. Operating current is approximately 1 A and suitable heatsinking is required as the junction temperature should be kept below 100°C. These LEDs are suitable for general illumination applications and are rapidly being adopted where long lamp life warrants the extra cost associated with LED technology compared to incandescent or fluorescent lamps.

A photograph of such a high-power LED is shown in Figure 5.28. Lumen maintenance is typically 70% of initial luminance after operating for 50,000 hours at a forward current of 1.0 A provided that the device is attached to a suitable heatsink. Based on GaInN material, a white colour is obtained by down-conversion using a phosphor material.

Visible LEDs are specified by radiometric units of optical power and wavelength of emission. In addition they are specified by luminous intensity and the *x* and *y* colour
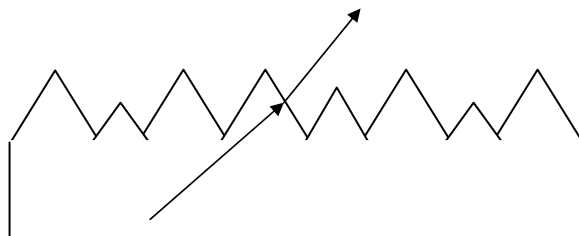
Figure 5.26 *Light propagation through a textured surface. The light path shown exceeds the critical angle of the surface but the beam can pass through due to surface texturing. Up to a 50% improvement in outcoupling has been achieved through the use of texturing*
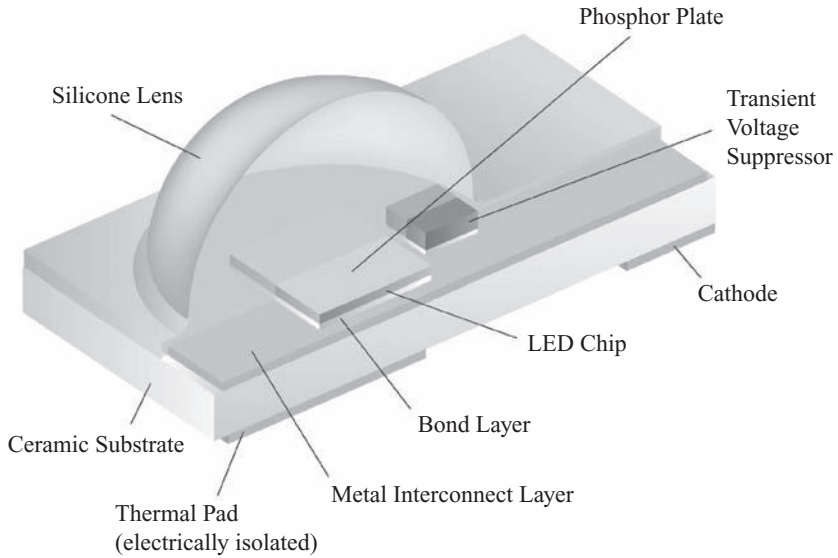
Figure 5.27 *Structure of high-power LED showing the die mounted on a base suitable for mounting on a heatsink. The lens is made of a silicone polymer, which withstands higher optical flux without yellowing compared to epoxy lenses. Reproduced from www.luxeon.com. Copyright (2011) LUXEON*

Figure 5.28 *Photograph of high-power LED capable of over 300 lumens output. The rectangular ceramic plate is designed for ease of mounting and heatsinking. Specific applications of these LEDs include street lighting, retail lighting and automobile headlights. Reproduced from www.luxeon.com. Copyright (2011) LUXEON*

coordinate values (see Chapter 3). Binning the LED die into separate groups is used to ensure a specific intensity range and colour space range for a given group of LEDs. This requires testing of each LED die and its selection based on spectral and intensity measurements.

In order to qualify as an illumination source the colour rendering index as well as the resulting colour coordinates of white-emitting LEDs must be carefully controlled. White LEDs can vary in colour from blue-white to yellow-white, and their colour coordinates will locate them the centre of the chromaticity diagram due to their broad emission spectra. Monochromatic LEDs having their emission peaks at wavelengths indicated by the scale marking the boundary of the CIE diagram will have x and y colour coordinates near this boundary close to relevant wavelengths (see Figure 3.13). LEDs having the narrowest emission spectra will be located nearest the boundary.

5.12   Summary

5.1. Efficient emission in all colours is available in LEDs. LED production is rapidly increasing to supply a variety of lighting markets. The combination of high efficiency, small size, long life and the absence of mercury used in fluorescent lamps are all driving forces behind this revolution.

5.2. Operation of LEDs is by means of electron-hole pair recombination. Photons generated must be outcoupled from the device and outcoupling losses should be minimized. Further losses occur due to non-radiative recombination.

5.3. The measured emission spectra of LEDs confirm the model presented in Chapter 3 for direct-gap electron-hole pair recombination.

5.4. Non-radiative recombination can take place at surfaces, interstitials, vacancies and anti-sites. Auger recombination is another possible recombination mechanism and is also non-radiative.

5.5. Optical outcoupling of a LED can be modelled by using Snell's law for light paths intersecting the semiconductor surfaces. Total internal reflection gives rise to an escape cone for light emerging through a planar surface of the LED die, and a lambertian emission pattern is typically present from the top surface of the die.

5.6. Historically GaAs LEDs were introduced in the 1960s. They provide infrared emission and take advantage of the liquid-phase epitaxy growth method and the direct bandgap of GaAs.

5.7. The interest in visible light emission and the incorporation of phosphorus into LEDs in the 1960s resulted in $GaAs_{1-x}P_x$ LED technology. As x increases in $GaAs_{1-x}P_x$ the bandgap increases. Visible emission starts at 750 nm, which corresponds to a bandgap of approximately 1.65 eV. Doping with nitrogen helps to promote radiative recombination and allows indirect gap materials such as GaP to exhibit sufficiently efficient luminescence for medium brightness green LEDs.

5.8. In the 1980s the $Al_xGa_{1-x}As$ system was developed in which the lattice constant is virtually independent of the aluminium content. Bandgap engineering opens the door to the double heterostructure, which provides carrier confinement and higher radiative recombination rates. In addition optical outcoupling is significantly improved. The

range of current density that a well can accommodate may be determined using a calculation based on a well density of states and a recombination coefficient.

5.9. AlGaInP LEDs developed in the late 1980s and early 1990s enables improved efficiency and a wider range of colours. The use of a quaternary system provides independent control of lattice constant and bandgap and shorter wavelength direct-gap emission is enabled. Luminous efficiency improves because of increased eye sensitivity as wavelength decreases. A transparent substrate may be wafer bonded to increase efficiency. MOCVD growth technique is versatile and allows a wide range of ternary and quaternary LED materials to be grown.

5.10. $Ga_{1-x}In_xN$ developed by the early 1990s has allowed high-efficiency high-brightness LEDs to cover the entire visible and the near-UV wavelength ranges. Development of nitride LEDs required new high-temperature growth conditions and the use of lattice-mismatched substrates and p-type doping with magnesium. Although double heterostructures are effective in $Ga_{1-x}In_xN$ LEDs, polarization favours the use of quantum wells, which provide better carrier confinement. Alternative crystallographic orientation can reduce or eliminate polarization. White emission may be achieved by down-converting blue emission using YAG phosphor.

5.11. LED optical outcoupling can be improved by fabricating dies with tilted sidewalls to reduce total internal reflections. Surface texturing can also improve outcoupling. Specially engineered materials and packaging are required to accommodate the increased cooling requirement and higher optical fluxes of power LEDs for lighting. Emission colour and colour rendering index are used to judge the white emission quality of a LED designed for general illumination purposes.

Suggestions for Further Reading

Nakamura S, Pearton S and Fasol G. *The Blue Laser Diode, The Complete Story*, 2nd edn. Springer, 2000.

Schubert E.F. *Light Emitting Diodes*, 2nd edn. Cambridge University Press, 2006.

Problems

5.1 Consider the LED emission spectrum and answer the following:
- (a) Obtain the full width at half maximum (FWHM) for the emission spectrum of the red LED of Figure 5.5 at room temperature.
- (b) Find the bandgap of the semiconductor used for the red LED of Figure 5.5. *Hint:* Treating E_g as a parameter, plot $R(E)$ vs E using Equation 5.1 and find the value of E_g that allows $R(E)$ to best match the spectrum of (a).
- (c) Using the value of E_g obtained from (b), plot Equation 5.1 at $-50°C$, $+20°C$ and $+100°C$ on the same graph and measure the FWHM for each plot.
- (d) Calculate the FWHM for the three temperatures of (c) using FWHM $= 1.8kT$ and compare with the answers of (c).
- (e) Is the photon emission rate $R(E)$ really equivalent to luminous intensity as a function of E? There are two sources of discrepancy, one based on photon properties and one based on photometric concepts. Explain both. Now explain why we are justified in neglecting these for the answers in (a) to (d).

5.2 The emission spectrum broadening discussed in Problem 5.1 is not the only possible broadening mechanism in a LED. List a few other additional contributions to linewidth broadening for the spectrum of a LED that may occur.

5.3 In Figure 5.17 the emission spectrum of a green $Ga_{1-x}In_xN$ LED is shown.
- (a) Obtain the full width at half maximum (FWHM) for the emission spectrum of the green LED at room temperature.
- (b) Estimate the bandgap of the semiconductor used for the green LED of Figure 5.17. *Hint:* Treating E_g as a parameter, plot $R(E)$ vs E using Equation 5.1 and find the value of E_g that allows $R(E)$ to best match the spectrum of (a).
- (c) Compare the answer to (a) with the predicted FWHM $= 1.8kT$.
- (d) What additional broadening mechanisms need to be considered in the emission spectrum of a $Ga_{1-x}In_xN$ LED?

5.4 Figure 5.24 shows optical outcoupling from a LED die with vertical sidewalls.
- (a) If all six sidewalls of the die are considered, what is the maximum outcoupling efficiency of a LED composed of GaAs in air?
- (b) If all six sidewalls of the die are considered, what is the maximum outcoupling efficiency of a LED composed of GaN in air?
- (c) Repeat (a) and (b) if the die is embedded in an epoxy resin with index of refraction $n = 1.5$.
- (d) There has been considerable effort to develop epoxy and silicone resins for LED packaging having higher indices of refraction. Repeat (c) if $n = 1.65$ and compare with your earlier answer to (c).
- (e) It is not normally practical to be able to use all six sidewalls of a LED die for light emission. Explain in the context of the packaged LED device in Figure 5.3, and revise your answers to (c) and (d).

5.5 An *integrating sphere* is commonly used to determine the total light output in lumens of a LED. It gathers all the light emitted from a LED regardless of the angle of emission of the light, and a detector then captures light with equal weighting from

all emission angles. In this way the emission angle does not influence the relative contribution to the total light measurement from the LED.

(a) Find more information on integrating spheres and show both a diagram and a photograph of an integrating sphere. Explain the operation of the integrating sphere in detail.

(b) The total light output in lumens of a LED may be obtained using an integrating sphere; however, it may also be obtained by determining the radiation pattern for the LED and then calculating the integral of this radiation pattern. For the radiation pattern of Figure 5.9, find the total LED output in lumens if the peak intensity of the LED is 10 000 candelas.

(c) For a given LED die placed in a package of the type shown in Figure 5.3, the radiation pattern will be determined by the shape of the epoxy lens and the geometry of the reflector cup. In Figure 5.9 a radiation pattern from a package yielding a 30° beam divergence is shown. If the same die is placed in a package with a 15° beam divergence, by what factor should the peak intensity of the LED increase?

(d) In commercial LEDs the peak intensity does not generally follow the theoretical increase in peak intensity when the package is changed from a 30° beam divergence to a 15° radiation pattern. Find the specification sheets for LEDs from a LED manufacturer that offers the option of 15° and 30° packages using the same LED die and show the manufacturer's specifications for both the 15° and 30° parts. Also show the peak emission intensity for both parts. By what factor does the peak emission increase in the 15° part?

(e) Explain the discrepancy in the peak emission ratios of (c) and (d). If total light output in lumens is to be maximized, which package would be best, a 30° beam divergence package or a 15° beam divergence package?

5.6 For the following direct-gap semiconductors, find and graph the ideal photon emission rate as a function of wavelength at room temperature if a LED were made from each material.
(a) GaAs
(b) GaN
(c) CdS

5.7 Look up more detailed information on liquid phase epitaxy (LPE). Explain the principles of the technique using a phase diagram and discuss which semiconductor types have been successfully grown using LPE. Why is LPE suitable for GaAs LEDs but not for GaN devices?

5.8 LED pioneer Nick Holonyak is credited with being the 'father of the LED'. Find and summarize a short biography of his life and discuss his contributions to LEDs.

5.9 Shuji Nakamura is a famous pioneer in the field of GaN LEDs. Find and summarize a short biography of his life and explain his contributions to the development of GaInN LEDs.

5.10 Look up more detailed information on molecular beam epitaxy (MBE). Explain the principles of the technique using a diagram and discuss which semiconductor types have been successfully grown using MBE. What are the important advantages and disadvantages of MBE?

5.11 Research more detailed information on the semiconductor growth technique metal-organic chemical vapour deposition (MOCVD). Explain the principles of the technique using a diagram and discuss which semiconductor types have been successfully grown using MOCVD. Why is MOCVD of particular importance in growing the more complex quaternary AlGaInP LED devices? Also show how it is used to grow GaInN LEDs.

5.12 The substrate materials used for GaN-based LEDs do not have ideal lattice constants or matched values of thermal coefficient of expansion to GaN. Find both lattice mismatch and thermal coefficient of expansion (TCE) mismatch for the leading substrate types Al_2O_3 and SiC relative to GaN. Show how the GaN lattice accommodates to these substrates.

5.13 The desire for affordable and high-quality single-crystal GaN wafers is an ongoing issue. Look into the following issues:
   (a) What is the fundamental difficulty in growing large single-crystal boules of GaN?
   (b) What methods have been developed to grow bulk GaN material?
   (c) For what applications is single-crystal GaN seen as important? Does this include LEDs?

5.14 The growth of GaN is generally based on *c*-axis growth on silicon carbide and sapphire substrates; however, the growth of GaN has been achieved in other crystallographic directions yielding non-polar M-planes or A-planes as well as semi-polar planes. Find more information about work done on these alternative growth directions and measured LED performance achieved. What specific challenges are limiting the widespread application of these alternative growth directions?

5.15 An $Al_xGa_{1-x}As$ double heterostructure LED without forward bias has the following parameters:
   Inside the well $E_f - E_c = 20\,\mathrm{meV}$
   Barrier height $\Delta E_c = 180\,\mathrm{meV}$
   Effective density of states $N_c = 5 \times 10^{17}\,\mathrm{cm}^{-3}$
   Carrier lifetime $\tau = 5 \times 10^{-9}\,\mathrm{s}$
   Carrier mobility $\mu_n = 1300\,\mathrm{cm^2\,Vs^{-1}}$
   (a) Calculate the room temperature leakage current density.
   (b) If the LED had a forward current density of $100\,\mathrm{\mathring{A}\,cm^{-2}}$, what fraction of total LED current does the leakage current represent?
   (c) Look up some LED data sheets for AlGaInP LED dies. Based on the manufacturer's specifications for a typical LED die, estimate the maximum forward current density in the die.
   (d) Draw a band diagram of a double heterostructure and explain the concept of leakage current in this context.

5.16 At 300 K at the electron well overflows in an $Al_xGa_{1-x}As$ double heterostructure with the following parameters:
   Barrier height $\Delta E_c = 180\,\mathrm{meV}$
   Well width $W = 150\,\mathrm{\mathring{A}}$
   Effective density of states $N_c = 5 \times 10^{17}\,\mathrm{cm}^{-3}$
   Recombination coefficient $B = 1.9 \times 10^{-10}\,\mathrm{cm}^{-3}\,\mathrm{s}^{-1}$
   (a) Calculate the current level at which the well overflows.
   (b) How does this overflow relate to the leakage current of Problem 5.15?

5.17 The distributed Bragg grating used in some LEDs is formed from a stack of alternating semiconductor layers having two alternating indices of refraction (see Figure 5.4). Show that for waves whose wavelength is four times the optical thickness of the layers, the many reflections combine with constructive interference, and the layers act as a reflector.

5.18 The growth of GaN on sapphire appears to have a very large lattice mismatch of 12%; however, a rotation of the axes in the growth plane of the GaN relative to the sapphire occurs that reduces the effect of this mismatch substantially. Look up more information on this topic. *Hint:* Use keywords in-plane rotation; GaN on sapphire growth.

    (a) Construct a diagram showing the planar arrangement of atoms in the sapphire substrate as well as the rotated GaN crystal plane that provides an optimum fit to the sapphire in the same diagram using two different colours to distinguish the layers.

    (b) What rotation angles are optimum?

5.19 Show that a lambertian source is obtained from the LED die of Figure 5.8. *Hint:* Find the area of a ring-shaped portion of a sphere mapped out in spherical polar coordinates in air with angles $d\Theta$ about Θ that corresponds to an area of a ring-shaped portion of a sphere inside the semiconductor having angles $d\theta$ about θ, and assume that θ is small.

    Now show that the luminous intensity of the source falls off with $\cos\Theta$ dependence as a function of Θ.

6

Organic Semiconductors, OLEDs and Solar Cells

Objectives

1. Understand conjugated polymers as distinct from polymers with saturated bonding.
2. Understand the mechanism of electric conductivity in conjugated polymers.
3. Explain absorption and emission processes as based on the molecular exciton.
4. Introduce polymer OLED devices and their structures.
5. Explain polymer OLED operation with reference to HOMO and LUMO levels.
6. Discuss materials used for polymer OLED devices and the band model of the polymer OLED.
7. Introduce small molecule OLEDs and their device structures.
8. Describe anode and cathode materials for small molecule OLEDs.
9. Discuss hole injection and transport layers as well as electron injection and transport layers in small molecule OLEDs.
10. Describe light emission in small molecule OLEDs including host materials and fluorescent and phosphorescent dopants.
11. Introduce the basic organic solar cell structure and discuss carrier dissociation at interfaces as well as the band model for the heterojunction solar cell.
12. Describe the advantages of more advanced organic solar cell structures including the bulk heterojunction and examples of self-organization.

6.1   Introduction to Organic Electronics

Organic materials have the potential to significantly change the way that electronic devices, including solar cells and LEDs, are manufactured. In the twentieth century the plastics revolution enabled low cost manufacturing of lightweight products such as plastic film for everything from food wrapping to thermal insulation for dwellings. The basic attributes of plastic that drove this included its formability into large-area sheets and the ability of the material to bend without cracking. Plastic film can be applied as a coating to a substrate or it may be manufactured directly as a film on a roll. It can be layered to give it desired properties such as moisture resistance, mechanical strength, thermal insulation and tolerance to a wide temperature range.

The field of *organic electronics* is much more recent. Here organic materials having well-defined electrical properties are being developed for electrical conductors, semiconductor devices and optoelectronic devices such as solar cells and LEDs. The replacement of brittle inorganic III-V semiconductors with organic materials for LEDs could open the door to large-area low-cost flexible lighting products or large-area flexible display screens. Likewise the replacement of brittle inorganic silicon solar cells with organic materials could enable large-area low-cost flexible solar cell sheets.

A few decades of research and development on organic electronic materials have taken place and a growing number of products that rely on such materials are in commercial production. Currently inorganic semiconductors still dominate the electronics industry. Nevertheless the inherent attributes of the organic materials just described continue to drive

their further development. This chapter reviews organic electronic materials and introduces their application to both light emitting devices and solar cells. Both the successes and ongoing challenges associated with organic materials will be discussed.

6.2 Conjugated Systems

Organic materials generally can be classified into *small molecule* organic materials and *large molecule* organic materials. Small molecule materials, often referred to as *oligomers,* contain only a few repeat units, or *mers*, per molecule, whereas large molecule materials, referred to as polymers, contain many mers per molecule. Typical polymer molecules contain hundreds or thousands of mers per molecule.

Electrically conductive organic molecules as electronic materials were understood in the 1970s, and they now constitute a new family of organic semiconductors that have the essential properties needed to make electronic devices. Conductive oligomers/polymers, however, have electrical properties that are orders of magnitude different from typical inorganic semiconductors and they may be processed at much lower temperatures. They can be used to make solar cells and light emitting devices as well as transistors and other electronic devices. We will begin by focusing on polymer materials and then in Sections 6.4–6.14 we will discuss small molecule materials followed by organic solar cells is Sections 6.15 and 6.16.

In order to understand the properties of these materials we need to review their molecular structure and understand the origin of the energy levels for electrons and holes as well as the mechanisms for electron and hole transport. We will compare these to inorganic materials to highlight the distinctive properties that must be considered in order to make use of conjugated polymers in semiconductor devices.

Conventional polymers are *saturated*, which means that the valence electrons of carbon atoms in a carbon chain are fully utilized in bonding and four atoms are bonded to each carbon atom. The simplest example of a saturated polymer is polyethylene, as shown in Figure 6.1. In polyethylene a carbon atom has six electrons, which normally occupy energy states $1s^2 2s^2 2p^2$.

For the carbon incorporated in polyethylene the four electrons in the $2s^2 2p^2$ shell achieve a lowest energy state in a *hybridized* sp^3 configuration that combines the spatial character of the s-orbital and the p-orbital resulting in the well-known tetrahedrally symmetric bonding
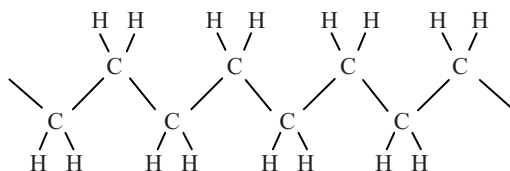
Figure 6.1   *The molecular structure of polyethylene. Each carbon has four nearest neighbours and forms four bonds. Polyethylene is an insulator and has a wide energy gap in the ultraviolet energy range. Each carbon atom has an almost perfect tetrahedral bond symmetry even though it bonds to both carbon and hydrogen nearest neighbours*
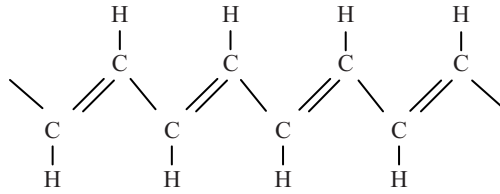
Figure 6.2 *Polyacetylene is the simplest conjugated molecule. It is often thought of as a chain of single bonds alternating with double bonds although actually all the bonds are equal and are neither purely single nor purely double. It consists of a carbon chain with one hydrogen atom per carbon atom. Since only three of the four valence electrons of carbon are used for bonding, one π electron per carbon atom is available for electrical conduction and becomes delocalized along the carbon chain*

character of carbon. This is demonstrated in inorganic diamond crystals but is also relevant to polyethylene. The bonding symmetry of a carbon atom in polyethylene is very close to tetrahedral, and the carbon bonds are *stereospecific*, meaning that the carbon bond angles are quite rigid and approach the tetrahedral bond angle of approximately 109°. It is interesting to note the relevance to inorganic silicon: silicon atoms, like carbon, exhibit similar hybridized, tetrahedrally symmetric bonding in crystalline silicon.

The simplest conductive polymer is polyacetylene. Its structure is shown in Figure 6.2. In this polymer there are three nearest neighbours leading to a different type of hybridization, which is denoted sp^2p_z. In this hybridization there is a lone electron called the π electron and three other electrons. These other three electrons take part in the bonding with the three nearest neighbours but the π electron does not and is left 'dangling' at each carbon atom. Polymers having this configuration are referred to as *conjugated*.

Since a chain of closely spaced carbon atoms is created, there are as many π electrons as carbon atoms in the chain, and there is a spatial overlap of these individual π electrons. The π electron at each atom interacts with all the other π electrons in the chain to form a number of energy bands. These bands are collectively referred to as *π-sub-bands*. Electrical conductivity also arises as a result of these π electron sub-bands since the electrons they contain become delocalized. As with inorganic semiconductors the Pauli exclusion principle forces the π electrons to occupy a range of energy levels leading to the formation of bands.

In addition we can regard such a conjugated molecule as a periodic potential in which the π electrons exist, which means that the Kronig–Penney model would predict energy bands and energy gaps. The length scale of one period in the periodic potential corresponds to the length of one mer on the polymer chain. Figure 6.3 shows a band diagram of a molecular solid containing a few molecules. The small potential energy barriers represent the separation between atoms or mers within a molecule, and the large barriers represent the separation between molecules. Intramolecular bonds are normally covalent or conjugated bonds, and intermolecular bonds are much weaker van der Waals bonds.

The fundamental band theory physics presented in Chapter 1 is relevant to molecules as well as inorganic materials: in addition to the periodicity of the molecule repeat units modelled by the periodic potential $V(x)$ in Equation 1.1, the overall dimension of the conjugated system, or *conjugation length*, is modelled as the length L of the quantum box
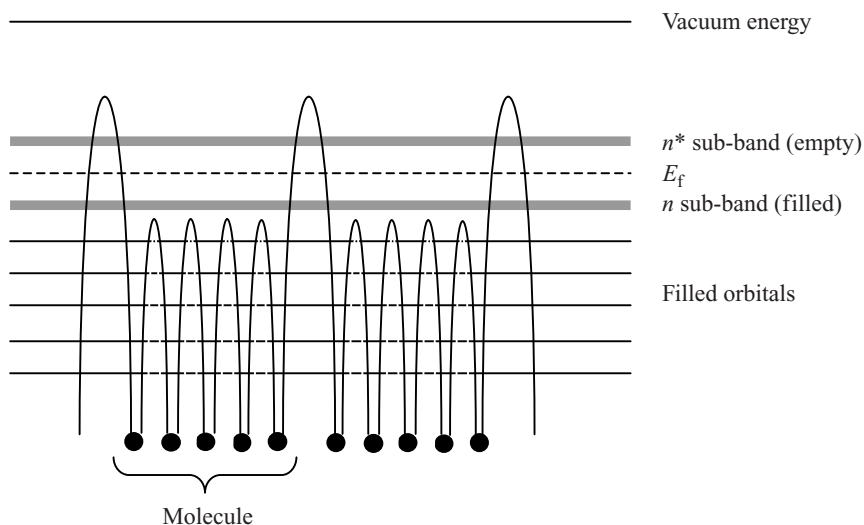
Figure 6.3 *Energy levels and bands in a few closely spaced organic molecules. Note the small energy barriers caused by the intramolecular bonding and the larger energy barriers caused by the intermolecular bonding*

of Example 1.1. The conjugation length refers to the effective length of the molecule over which electron delocalization occurs. This conjugation length acts as the length L of a one-dimensional box in which the electron is confined, and the simplest model of the molecule is therefore analogous to Example 1.1. The longer the conjugation length the smaller the energy level spacing. More complete modelling based on this foundation can then predict molecular energy levels, but we will not undertake this in this book.

The number of electrons relative to the number of available states in the π sub-bands determines the electrical nature of the polymer. If the π sub-bands are either filled or empty then the material will be semiconducting and if the highest filled band is partly filled then the material will be metallic. The energy difference between the highest occupied π sub-band and the lowest unoccupied π sub-band determines the energy gap in the case of a semiconducting polymer. The unoccupied π bands are commonly referred to as π^* bands.

A wide range of large molecules exist that have semiconducting properties. A number of these are shown in Figure 6.4. Note that with the exception of polyacetylene, the remaining molecules have ring elements containing five or six atoms per ring. The simplest ring element is the benzene ring having six carbon atoms and six delocalized π electrons, one from each atom. Six-sided rings are also the building blocks of graphite, a conductive material. It is therefore not surprising that ring-containing molecules are prominent among conducting polymers.

The macroscopic electrical conductivity of a polymer material depends both upon the conduction within a molecule (*intramolecular* conduction) as well as the ability for charges to transfer from molecule to molecule (*intermolecular* conduction). Intermolecular conduction is strongly dependent on the proximity of molecules to each other and their orientation

Polyacetylene
(PA)

Poly para-phenylene
vinylene
(PPV)

Polythiophene
(PT)

Poly(2,5-dialkoxy)
paraphenylene
vinylene
(derivative of MEH-PPV)

Poly(3-alkyl)
thiophene
(P3AT)
(R-methyl, butyl, etc.)

Poly para-phenylene
(PPP)

Polypyrrole
(PPy)

Poly
isothianaphthene
(PITN)

Ladder-type
polyparaphenylene
(LPPP)

Poly ethylene
dioxythiophene
(PEDOT)

Polyparaphenylene
sulphide
(PPS)

Polyheptadiyne
(PHT)

Alkoxy-substituted
poly *para*-phenylene vinylene
(MEH-PPV)

Poly(3-hexyl) thiophene
(P3HT)

Polyaniline (PANI)

Polyfluorene (PFO)

Figure 6.4*   Molecular structures of well-known conjugated polymers. Many molecules con-
tain a combination of linear and ring-type structures, the simplest example being poly para-
phenylene vinylene (PPV). Reprinted from Li, Z., Meng, H., Organic Light-Emitting Materials
and Devices, 157444-574X. Copyright (2006) with permission from Taylor & Francis*

*Full molecular names such as Poly(2,5-dialkoxy) paraphenylene vinylene will not generally be listed in
this chapter but may be found in Suggestions For Further Reading, Z. Li et al.
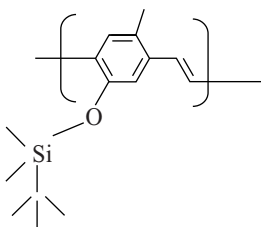
Figure 6.5   *Poly para-phenylene vinylene (PPV) derivative forming a silicon-substituted soluble polymer*

to each other. Crystalline polymers are generally more densely packed; however, both amorphous and crystalline phases are candidate materials; crystalline polymers often contain crystalline domains in an amorphous matrix.

Carrier mobility values for organic semiconductors typically span the range from $10^{-7}\,\mathrm{cm^2V^{-1}s^{-1}}$ to about $1\,\mathrm{cm^2V^{-1}s^{-1}}$. Note that these values are much lower than a typical inorganic semiconductor like silicon with mobility on the order of $10^3\,\mathrm{cm^2V^{-1}s^{-1}}$.

There are a number of interesting opportunities offered by molecular semiconductors:

(a) *Chemical properties* can be modified by changing the side-groups on the molecules. For example, in Figure 6.4, PPV is insoluble in organic solvents. Since PPV contains a benzene ring it can be modified by adding groups to the ring, and soluble molecules are obtained, which are referred to as *PPV-derivatives*. One example of a soluble PPV derivative is shown in Figure 6.5. A soluble PPV derivative allows a polymer with the electrical functionality of PPV to be formulated in a solution, deposited as a liquid and then baked to release the solvent. This is a very cost-effective device preparation technique that wastes very little material.

(b) *Electronic properties* can be modified by changing side-groups on the molecule. For example, PPV can be modified to become MEH-PPV shown in Figure 6.4 by adding the indicated groups to the benzene ring of PPV. Figure 6.6 shows PPV and derivatives that are frequently used in polymer OLEDs. This modifies the energy gap of PPV as shown allowing absorption or emission wavelengths to be tuned to desired values, which is of particular relevance to optoelectronic applications.

The absorption process excites an electron using a photon from the highest occupied π sub-band to the lowest unoccupied π^* sub-band; the emission process generates a photon when the electron falls back to its original band. An exciton is formed and then annihilated during emission as discussed in detail in Chapter 3. The *photoluminescence efficiency* of this absorption and emission process can be in the range of 10% to over 50%, which makes these materials suitable for high-efficiency optoelectronic device applications.

Electroluminescence efficiency is lower than photoluminescence efficiency due to spin associated with the hole and electron within the exciton. The maximum efficiency under electroluminescence is normally 25% of the photoluminescence efficiency, although efficiency can be higher if phosphorescence occurs. These optical processes were discussed in Section 3.7. Conjugated polymers are suitable candidates for both organic solar cells and organic light emitting diodes.
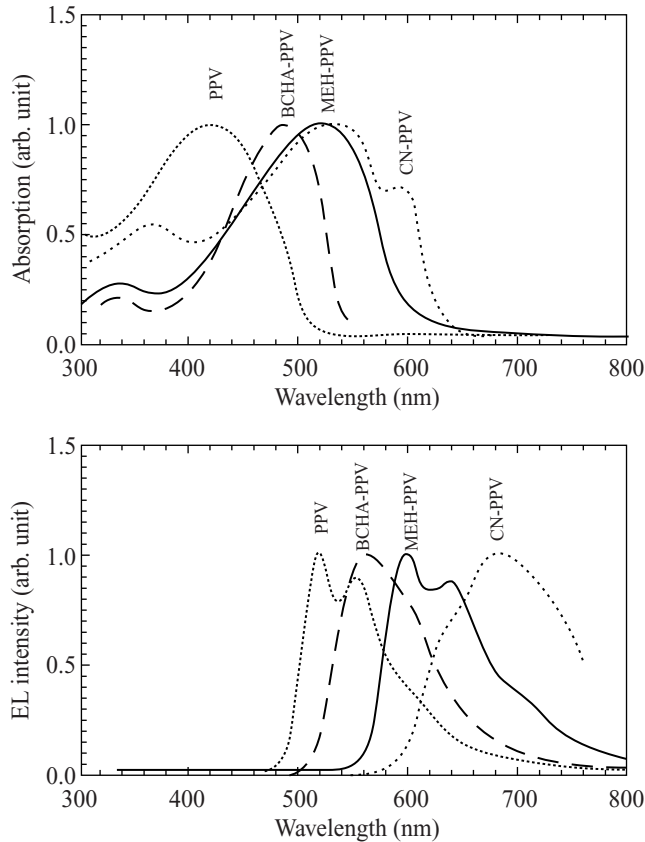
Figure 6.6   *Absorption and emission of poly para-phenylene vinylene (PPV) and PPV deriva-tives. The energy gap determines the upper wavelength range of absorption as well as the lower wavelength range of emission. Here energy gaps from 1.9 eV ($\cong$ 640 nm) to 2.5 eV ($\cong$ 500 nm) result in these spectra. Reprinted from Li, Z., Meng, H., Organic Light-Emitting Materials and Devices, 157444-574X. Copyright (2006) with permission from Taylor & Francis*

6.3   Polymer OLEDs

The structure of a basic polymer *organic light emitting diode* (OLED) is shown in Fig-ure 6.7. A glass substrate is coated with a transparent conductor such as indium tin oxide (ITO). The ITO must be as smooth as possible because the subsequent conjugated polymer layer is generally under 100 nm thick. An EL polymer layer is deposited on the anode layer and finally a low-workfunction cathode layer completes the device. This cathode layer is composed of group I or group II metals or compounds that are easily ionized.

Deposition techniques of *solution processing* are applicable to most electronic polymers. A solution containing the desired polymer can be formed as a liquid and spread over the substrate. The solvent can then be evaporated away leaving a layer of the desired polymer. This is a low-cost deposition technique and is a key advantage of polymer organic materials
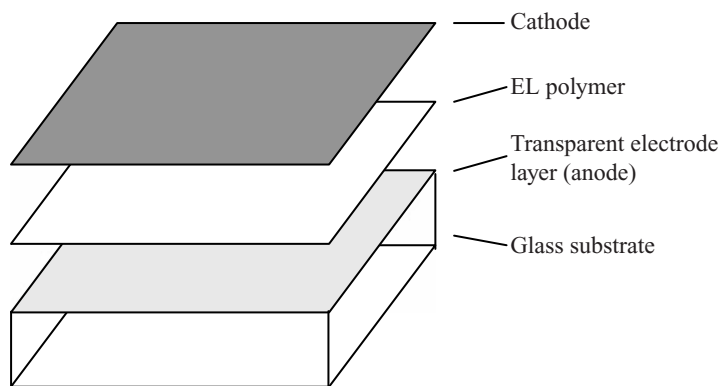
Figure 6.7 Structure of basic polymer OLED consisting of a glass substrate, a transparent ITO anode layer, an EL polymer layer and a low workfunction cathode layer

compared to the more expensive deposition methods required for small molecule materials discussed in Section 6.4.

The EL polymer layer provides three main roles in the polymer OLED:

(a) Near the cathode the EL polymer acts as an electron transport layer.
(b) Near the anode the EL polymer acts as a hole transport layer.
(c) The EL polymer provides for the recombination of holes and electrons via molecular exciton annihilation and the consequent emission of light.

We can understand the ability of one layer to serve all these functions by understanding the workfunction difference between the two electrodes. Assume a cathode and an anode material are placed in vacuum, are electrically isolated from each other and are not connected to any voltage source. The anode has a high workfunction, typically in the range of 5 eV, and the cathode has a low workfunction in the range of 1–2 eV. This means that for an electron to leave the surface of the anode and enter vacuum requires about 5 eV but an electron leaving the cathode requires only about 1–2 eV to enter the same vacuum. This may be illustrated by defining a vacuum energy E_{vac} for both anode and cathode as shown in Figure 6.8.

Since the electrons leaving the anode and cathode leave from their Fermi levels (highest occupied energy levels) it is clear that we cannot draw a diagram simultaneously showing aligned Fermi levels and a constant value of E_{vac}. This problem is resolved since an electric field is generated in the vacuum giving rise to a potential gradient between the anode and cathode that accounts for the difference in work function. A built-in charge is present on each electrode to create this electric field: the anode carries a *negative* charge and the cathode carries an equal and opposite *positive* charge. For the charges to build up, there must be an opportunity for charge to transfer between anode and cathode; however, once the appropriate charges are present there will be no further charge transfer and equilibrium is maintained. Without the opportunity to transfer charge equilibrium can never be reached.

We can now return to the case of the EL polymer layer sandwiched between anode and cathode. The low carrier concentration in the EL polymer layer ($\leq 10^{14}$ cm^{-2}) combined
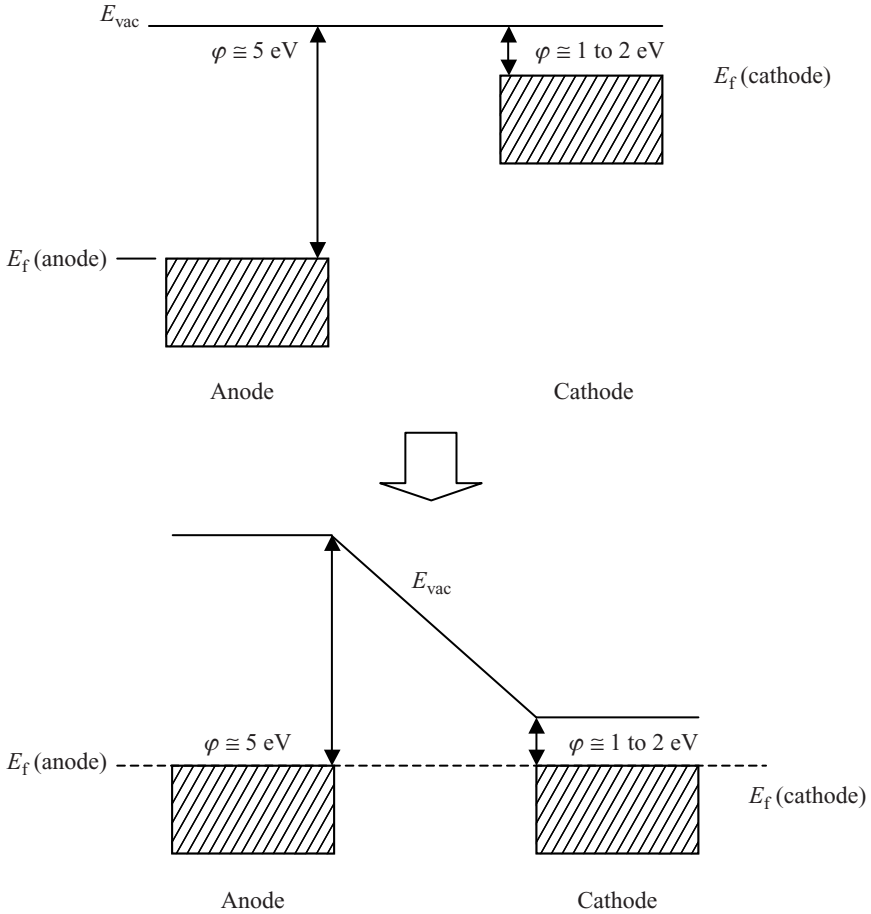
Figure 6.8 *Energy diagram showing a high workfunction anode and a low workfunction cathode. With a constant vacuum energy the Fermi levels cannot be aligned (upper diagram) and this diagram is therefore not an equilibrium diagram, which invalidates the concept of Fermi energy. To resolve this problem an electric field forms between anode and cathode and an equilibrium diagram (lower diagram) having aligned Fermi energies is the result. This is the result of charge transfer*

with its low carrier mobility means that we can regard this layer as a virtual insulator sandwiched between the anode and cathode electrodes. There will be a built-in electric field within the polymer caused by the workfunction difference between the two metals, as shown in an ideal form in Figure 6.9a, which is drawn for equilibrium conditions. A slope in the upper π^* band and the lower π band is present, which is a consequence of the electric field present in the EL polymer. The finite but small conductivity of the EL polymer is sufficient to transfer the necessary charge for equilibrium to be reached.

If a voltage equal to the difference between the two workfunctions is applied across the anode and cathode the *flat-band* condition is reached. The anode is biased positive with
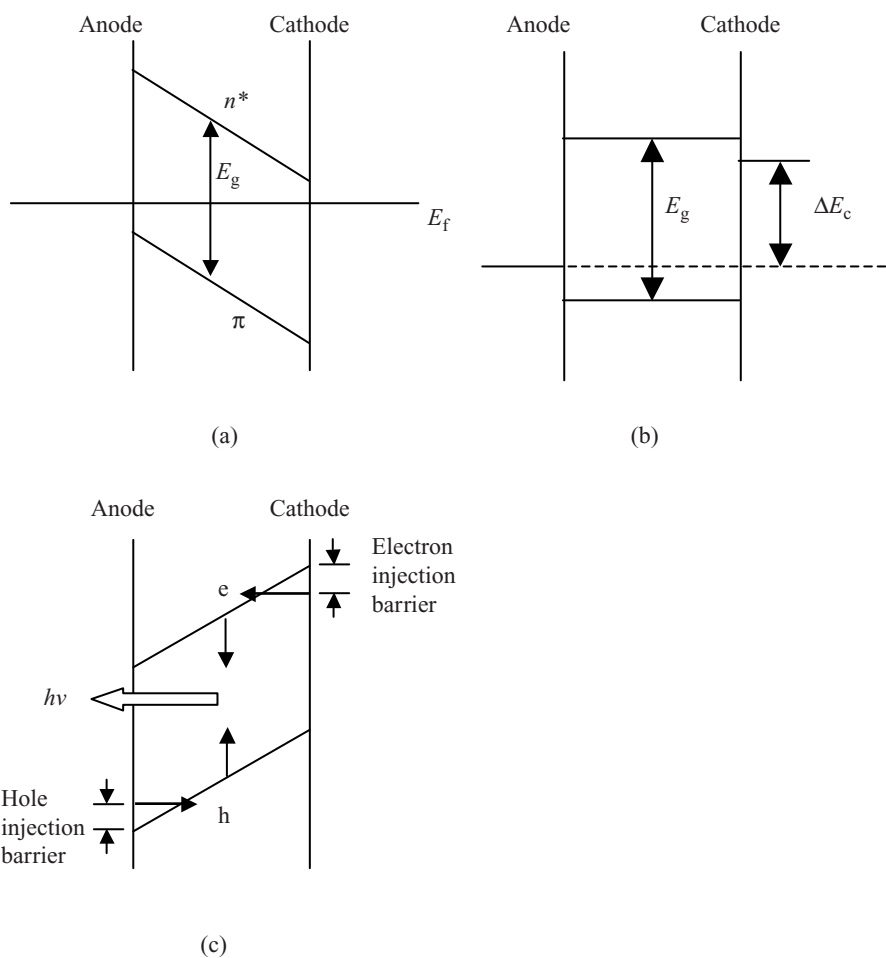
Figure 6.9 *The upper π* band and lower π band in a polymer EL layer. (a) The equilibrium condition. (b) The flat-band condition in which a positive voltage is applied to the anode. (c) Device in forward bias in which holes and electrons are injected and form molecular excitons, which annihilate to generate photons. The resulting hole injection barrier and electron injection barrier are shown*

respect to the cathode. At this voltage, which is referred to as the *flat-band voltage*, the EL polymer energy bands are horizontal and no electric field exists. The applied voltage causes additional charges to be added to the anode and cathode that cancel out the built-in charges. The flat-band condition is shown in Figure 6.9b. The difference between the conduction band edges of the two electrodes, ΔE_c, is shown.

The vacuum workfunctions of the anode and cathode materials are actually modified by the details of the electrode-EL polymer interfaces, which also affect the flat-band voltages. This modification of workfunctions is caused by trapped charges at the interfaces and other impurity and diffusion effects that occur there. A similar set of issues was described in

Section 2.9 for inorganic metal-semiconductor contacts, although the polymer EL material offers particularly high diffusivity even at room temperature and is therefore more prone to migration of undesired species.

If the anode voltage is made still more positive then the situation in Figure 6.9c results. Here the electrons from the cathode tunnel across the potential 'spike' at the cathode edge of the π^* band and are injected into the EL polymer. The electric field is in the opposite direction to that shown in Figure 6.9a, and electrons that enter the π^* band start drifting towards the anode. It is important to minimize the energy barrier seen by cathode electrons as they are injected into the band. The electrons will enter the lowest energy state in this π^* band and we refer to this as the *lowest unoccupied molecular orbital* (LUMO). Note that this is analogous to E_c, the bottom of the conduction band, in inorganic semiconductors. Generally a difference between cathode Fermi energy and the LUMO energy of 0.2 eV or less is suitable.

Proper modelling of electron current flowing across this potential barrier also requires the application of thermionic emission physics, which was discussed in detail in Section 2.9. In practice both thermionic emission and tunnelling are often involved.

At the same time, holes can be injected into the π band from the anode. It is more correct to describe this as the formation of holes at the anode-EL polymer interface due to the transfer of electrons from the π band to the anode. These holes are swept by the electric field towards the cathode in the highest available π band energy level, which is referred to as the *highest occupied molecular orbital* (HOMO), and the holes have an opportunity to combine with π^* band electrons to form excitons. Upon annihilation of these excitons, light is emitted.

Cathode metals require the use of low-ionization-energy elements, and examples of cathode materials include a thin barium-based, calcium-based or lithium-based layer in contact with the active polymer and then possibly coated with a final aluminium capping layer to offer some protection and to improve sheet conductivity. Since group I and group II metals are very easily oxidized and react vigorously with water, care must be taken to protect them from the atmosphere. OLED devices in the laboratory must be handled and tested in a very dry inert gas environment requiring the use of a glove box.

Anode materials are generally transparent to allow light to leave the active layers, and indium tin oxide (ITO) is popular. There are three challenges with ITO, however:

(a) ITO is a polycrystalline film and has naturally occurring surface roughness, which can be comparable in scale to the thickness of the EL polymer layer (20–100 nm). This can cause OLED failure due to short circuits.
(b) Indium is highly mobile in EL polymers and can migrate from the anode into the active region.
(c) The ITO surface workfunction is highly unstable and inhomogeneous. There are many dangling bonds on the ITO surface, which need to be hydroxide-terminated. The surface of the ITO contains Sn atoms that substitute for In atoms and add to the chemical complexity.

To prevent these difficulties an additional conductive polymer anode layer is commonly inserted between the ITO and the EL layer. This layer has a HOMO level close to the HOMO level of the EL layer to promote hole injection. It also acts as a planarization layer to create a smooth surface for the EL polymer layer. As a result of adding a polyaniline
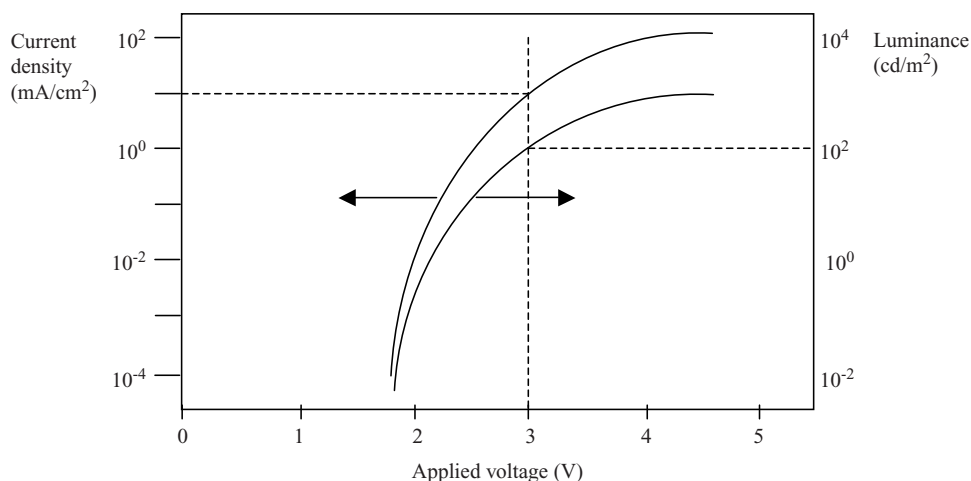
Figure 6.10 *Typical luminance–voltage (L–V) and I–V characteristics of a polymer OLED. A well-defined threshold voltage is observed due to the sharp onset of carrier injection from the electrodes across the potential barriers at the electrode-EL polymer interfaces. Note the similarity between the shapes of the current and luminance curves*

(PANI) anode layer (Figure 6.4) an increase in the life of a PPV-based polymer OLED from hundreds of hours to over 10 000 hours has been obtained.

Typical operating characteristics of a polymer OLED are illustrated in Figure 6.10. Both current and luminance are shown as a function of applied voltage. Below a threshold voltage of about 1.8 V, current does not flow and there is no light output. Both current and luminance rise rapidly as voltage is increased and carriers are able to be injected into the LUMO and HOMO levels. The luminance is essentially proportional to the current over several orders of magnitude.

The performance of OLEDs is frequently expressed in terms of $cd\,A^{-1}$. In Figure 6.10 at an applied voltage of 3 V the current density is $10\,mA\,cm^{-2} = 100\,A\,m^{-2}$ and the luminance is $100\,cd\,m^{-2}$. This means that the operating efficiency at 3 V is $(100\,cd\,m^{-2})/(100\,A\,m^{-2}) = 1.0\,cd\,A^{-1}$. Polymer OLEDs with operating efficiency values of well over $10\,cd\,A^{-1}$ have been achieved. In addition, very high peak luminance values of up to $\cong 10^6\,cd\,m^{-2}$ may be obtained if large transient currents in the range of $100\,A\,cm^{-2}$ are applied.

The most challenging aspect of OLED performance has been operating lifetime, which normally is defined as the operating time under constant current conditions for the OLED to decrease in light emission to half its initial intensity. Purity and interface stability are two requirements of practical OLED materials. It is found that lifetime can be well correlated to the total charge density $\rho_{accumulated}$ passing through the OLED during its lifetime. This means that lifetime can be determined for an OLED that varies with time in brightness and current density by integrating the current density over time until half intensity is reached, and

$$\rho_{accumulated} = \int_0^T J(t)\,dt$$

Typical values of $\rho_{\text{accumulated}} \geq 10^5 \, \text{C cm}^{-2}$ are achievable in a sealed, dry environment at room temperature.

6.4   Small-Molecule OLEDs

In addition to polymer EL materials, *small molecule* organic materials can also be used to fabricate efficient LED and solar cell devices. Examples of small molecules with useful optoelectronic properties are shown in Figure 6.11. Small-molecule materials are generally not soluble and are deposited by *vapour deposition* methods. Using a vacuum chamber, thin films of these materials can be grown by heating a source pellet or powder of the material causing it to evaporate onto a substrate. Using an appropriate chamber design deposited films with good thickness uniformity and high purity can be obtained; however, there is significant material waste due to the evaporation process. The small-molecule materials in Figure 6.11 all contain ring structures and have delocalized electrons for high intramolecular conduction. Intermolecular conduction is also critical for bulk conductivity to exist.

There are several requirements of the small-molecule materials:

(a) They must not decompose during thermal evaporation in the vapour deposition process. A source pellet of the material is heated to temperatures in the range from $150°C$ to over $400°C$ in a vacuum chamber to create a high enough vapour pressure for the molecules to evaporate and condense on a cool substrate positioned a set distance away from the source material.
(b) The deposited films must be of high quality and purity. Typical thicknesses are in the range from 5 to 200 nm.
(c) The glass transition temperature T_g should be high enough to prevent unwanted crystallization under the conditions of normal operation. Typically T_g should be over $85°C$.

Unlike polymer OLEDs, small-molecule devices typically use separate electron- and hole-transporting materials and are therefore more analogous to inorganic p-n junction devices. A basic structure of a small-molecule OLED is shown in Figure 6.12. A *hole transport layer* (HTL) and a separate *electron transport layer* (ETL) form a junction sandwiched between anode and cathode electrodes.

It is also possible to introduce a *light emitting material* (LEM) positioned between the HTL and ETL, which is optimized for radiative recombination efficiency and in which the holes and electrons form excitons. In these materials a number of additional energy transfer processes are often involved, which are described in more detail in Sections 6.11–6.14.

The finally process requires photon outcoupling. There are internal reflections that limit the escape cone of the photons. This is similar to the situation in inorganic LEDs; however, the refractive index *n* of the active organic layers is lower than for inorganic semiconductors, which allows better outcoupling. A typical value of $n \cong 1.5$ allows 20–30% outcoupling through the glass substrate.

The small-molecule OLED with separate HTL and ETL materials is more efficient than single-layer designs. This is true since each layer can be individually optimized for carrying one type of charge carrier. In addition, recombination near the interface of the two layers is
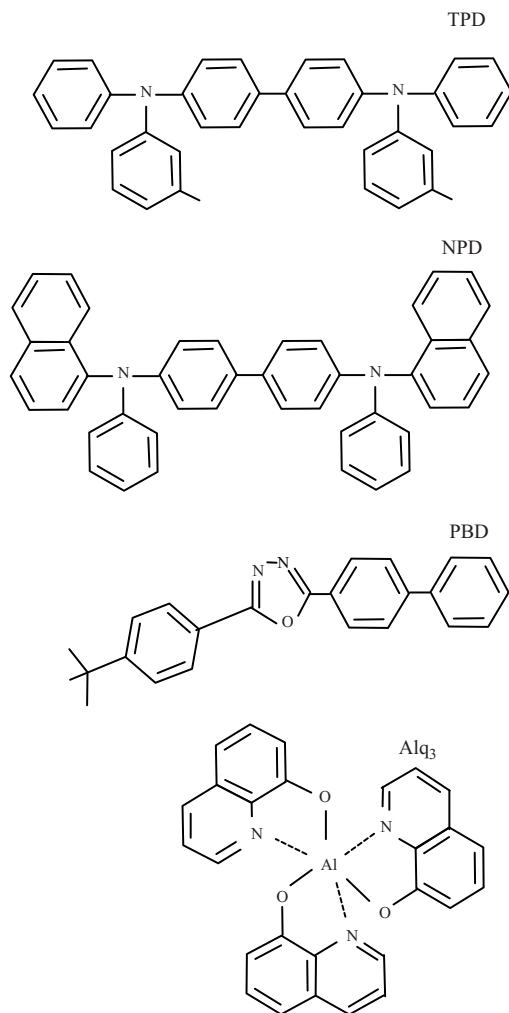
Figure 6.11 *Small organic molecules used for small-molecule OLED devices. Hole-transporting materials are TPD and NPD. Electron-transporting materials are PBD and Alq₃. Reprinted from Li, Z., Meng, H., Organic Light-Emitting Materials and Devices, 157444-574X. Copyright (2006) with permission from Taylor & Francis*

achieved. This prevents recombination from occurring too close to the anode and cathode layers, which has the effect of quenching luminescence and causing non-radiative decay of the excitons. Even better performance can be obtained by adding the light emitting material as well as an anode *hole injection layer* (HIL). The cathode is shown to include an *electron injection layer* (EIL), which is composed of LiF in Figure 6.12. An OLED structure containing these layers is shown in Figure 6.13.

The operation of the small-molecule OLED can be better understood with reference to the band diagram shown in Figure 6.14. Upon application of an electric field electrons

268 Principles of Solar Cells, LEDs and Diodes

Al
LiF
Electron transport layer (ETL)
Hole transport layer (HTL)
ITO anode
Glass substrate
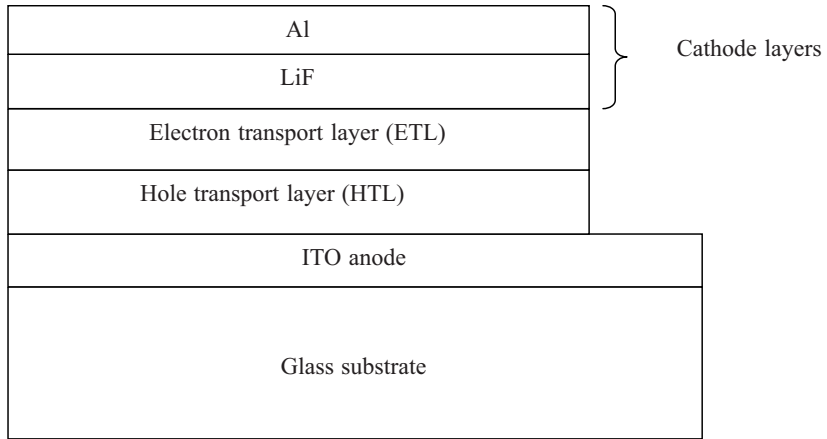
Cathode layers

Figure 6.12 *Small-molecule OLED structure. The OLED includes a transparent substrate, transparent ITO anode, hole transport layer (HTL), electron transport layer (HTL) and cathode. HTL materials such as TPD or NPD and electron transport materials such as Alq$_3$ or PBD are suitable. A popular cathode is a two-layer Al/LiF structure as shown*

Al
Electron injection layer (EIL)
Electron transport layer (ETL)
Light emitting material (LEM)
Hole transport layer (HTL)
Hole injection layer (HIL)
ITO anode
Glass substrate
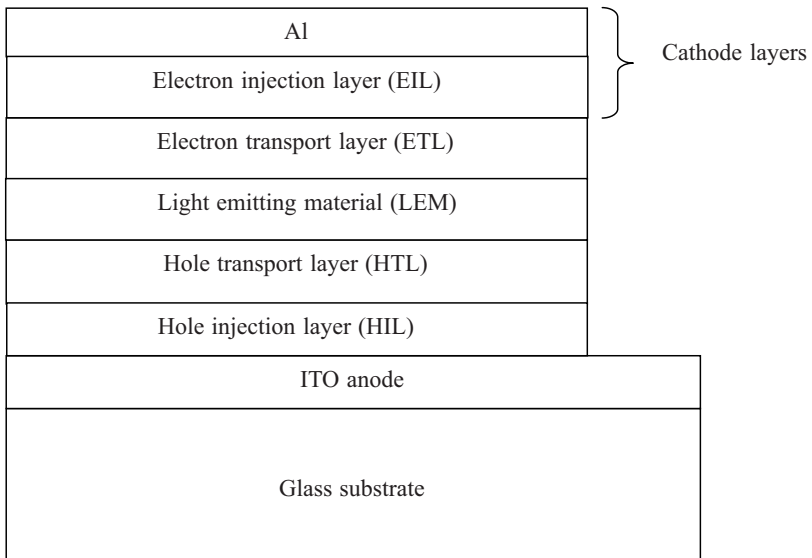
Cathode layers

Figure 6.13 *A more optimized small-molecule OLED structure includes an electron injection layer, a hole injection layer and a light emitting material. The cathode includes the electron injection layer*
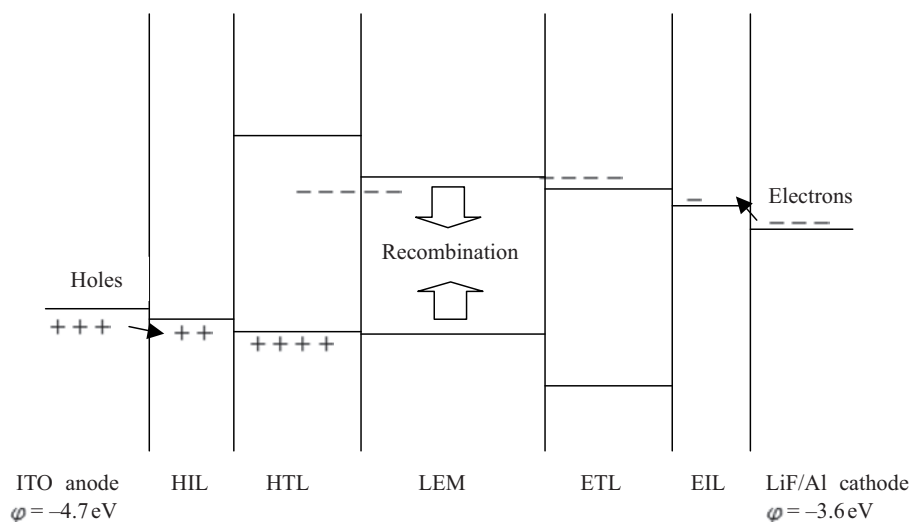
Figure 6.14 *Band diagram of a small-molecule OLED showing LUMO and HOMO levels for the various layers of the device. The band diagram is drawn without a bias applied. The accepted workfunctions of anode (ITO) and cathode (LiF/Al,) which are 4.7 eV and 3.6 eV respectively, are shown*

are injected by means of the EIL and ETL while holes are injected by means of the HIL and HTL. These holes and electrons meet in the LEM and form excitons there. Optimum performance is obtained when the same number of active holes and electrons enter the LEM per unit time, otherwise excess carriers of one type or the other are unable to participate in the recombination process, which causes efficiency loss. This is a current balance condition between holes and electrons similar to that discussed in the context of inorganic LEDs in Section 5.4.

In operation, holes are injected by means of the anode into the HTL and electrons are injected from the cathode into the ETL. With a suitable applied voltage a sufficient electric field exists for the holes and electrons to drift across their respective layers. The mobilities of these layers are important; compared to inorganic p-type and n-type materials mobility values are much lower. A typical hole mobility for small-molecule HTL materials is $10^{-3}\,\mathrm{cm^2V^{-1}s^{-1}}$, and a typical electron mobility for small-molecule ETL materials is 10^{-5} to $10^{-4}\,\mathrm{cm^2V^{-1}s^{-1}}$. These values are sufficient provided the layers are very thin; layer thicknesses in the range of only 10–200 nm are typical.

Next, exciton formation occurs by coulombic attraction of the holes and electrons as they encounter each other. This normally occurs within the transport layers and since the HTL typically has a higher mobility the ETL may be the layer in which excitons are formed. Photon emission from these excitons is normally limited to singlet excitons that decay very rapidly. The excited state has a very short spontaneous lifetime typically in the range of 10 ns due to the strong dipole strength of the radiating exciton. This means that the radiation originates very close in location to where the exciton was created.

There are a large number of small molecules that have been investigated for small-molecule OLEDs, and new molecules with specific attributes continue to be researched. Sections 6.5–6.14 present selected examples of molecular materials designed for various layers in small-molecule OLEDs as well as their most important properties.

6.5   Anode Materials

The ITO anode offers ease of patterning and good stability. As with polymer OLEDs, the surface smoothness of both the substrate and the ITO layer is important. A surface roughness below 2 nm is generally required. ITO has a high workfunction ($\varphi > 4.1\,\text{eV}$) allowing it to inject holes efficiently. The transparency of the ITO is a result of its wide bandgap of over 4 eV.

There are challenges associated with the use of ITO for small-molecule OLEDs, some of which were also noted for polymer OLEDs. ITO has a resistivity of $\cong 2 \times 10^{-4}\,\Omega\,\text{cm}$, which limits the current flow through an ITO layer and results in unwanted voltage drops along the anode conductors, which is particularly problematic if narrow anode rows or columns are required for an OLED display. See Problem 6.3. ITO is inevitably rough due to its polycrystallinity. It has a chemically active surface that can cause migration of indium into subsequent polymer layers. The ITO workfunction is sensitive to the cleaning process used to prepare the ITO for subsequent processing. ITO is a brittle inorganic film that is not ideal on polymer substrates because it normally requires high temperatures to deposit (200–400°C) and can crack due to mechanical and thermal deformation on polymers. ITO nevertheless remains the most common anode material for OLEDs.

6.6   Cathode Materials

Unlike anode materials, cathode materials are generally not transparent, which provides a wider range of materials choices. They must provide high conductivity and low workfunction, and good adhesion to the underlying polymer layers. Stability is also important and is highly dependent on packaging.

Challenges associated with cathodes include ease of oxidation, which is a consequence of the low-workfunction materials that require easily ionized group I and group II metals. There is also a tendency for these cathode layers to cause chemical reduction of adjacent organic layers. As with polymer OLEDs a two-layer cathode is popular, and the LiF/Al structure is widely used in which the aluminium protects the reactive LiF layer and also provides improved sheet conductivity. The formation of LiF preserves the low workfunction of the anode but reduces its tendency to oxidize as a result of reaction with the organic EIL material. LiF/Al cathodes have a workfunction of 3.6 to 3.8 eV.

The cathode layer is usually the most reactive layer in the OLED in the presence of oxygen or water. OLED devices must be protected with encapsulation. In practice glass and metal or glass sheets cover and seal OLED devices on the front and back respectively. An epoxy edge seal between front and back sheets is used; however, moisture can slowly diffuse through this polymer layer. Inside the OLED package a sacrificial *getter* material may be used to scavenge moisture or oxygen. An OLED package is shown in Figure 6.15.
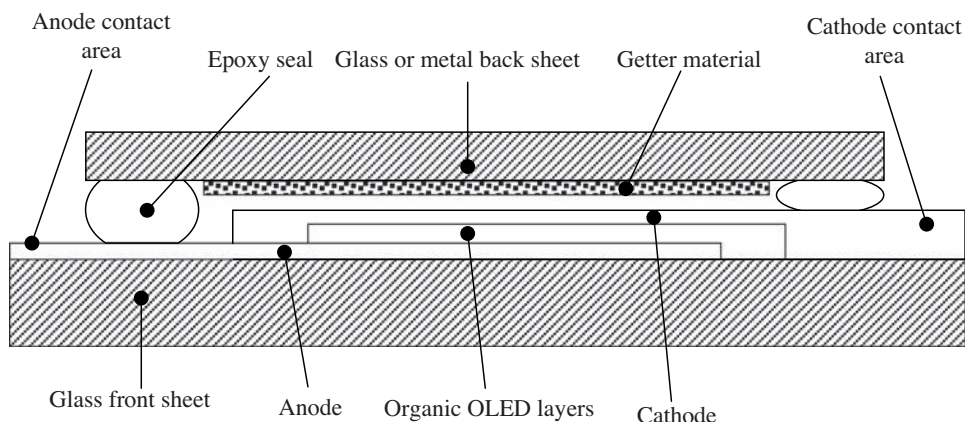
Figure 6.15   *OLED package includes front and back sheets, epoxy seal material on all edges, sacrificial dessicant or getter material, cathode and transparent anode having cathode contact areas for external connections. The rate of moisture penetration must be calculated to ensure a specified product life*

6.7   Hole Injection Layer

The HIL acts to improve the smoothness of the anode surface due to the native ITO roughness. It also facilitates efficient hole injection and long-term hole injection stability. The polymer-metal interface is very complex due to charges that are trapped there as a result of dangling bonds and metal atoms that react with the organic layer. Materials used for the HIL include copper phthalocynanine or CuPc, which is a widely used *porphyrinic metal complex*. Its molecular structure is shown in Figure 6.16.

This material improves efficiency and life in OLEDs; however, the precise mechanisms for the improvements are controversial. CuPc may reduce the hole injection barrier, which is equivalent to the hole injection barrier illustrated in Figure 6.9c for polymer OLEDs. In addition CuPc has a particularly good wetting characteristic on ITO, which may contribute to better interface strength. CuPc has good thermal stability and can reduce the dependence of the hole injection barrier on the cleaning procedure used on the ITO layer. CuPc is a semiconductor, which gives it sufficient electrical conductivity to carry current between the anode and the HTL.

Another important aspect of OLED performance relates to the current balance between electron and hole currents. CuPc may in some cases also improve this current balance and therefore the efficiency of OLED devices. This is highly dependent on the subsequent layers and particularly the electron injection and transport materials used in the OLED. Hole current generally tends to be in excess compared to electron current in OLEDs due to the higher hole mobility in organic materials. This suggests that CuPc may controllably decrease hole current in some cases. CuPc and phthalocynanines in general are *ambipolar*, meaning that they can act as both hole and electron conductors although the electron mobility is one to two orders of magnitude less than the hole mobility. Sometimes these layers are called *electron-blocking* layers.
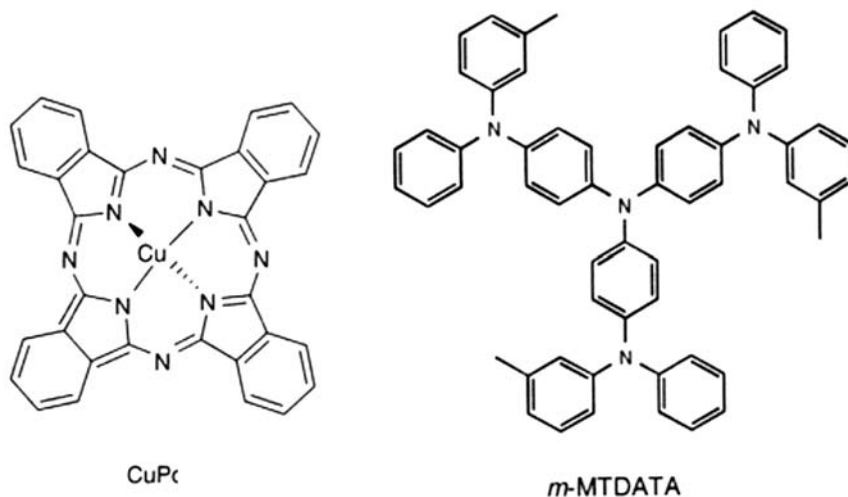
Figure 6.16   *Copper phthalocynanine, or CuPc, a widely used metal complex used for the HIL in small molecule OLEDs. Another HIL molecule is m-MTDATA, also shown. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

A challenge in the use of CuPc is the material's tendency to crystallize over time. The crystallization process is thermally driven. This may effectively contribute to limiting the ultimate device lifetime. Other materials that are useful in forming desirable ITO interfaces are under investigation and include fluorocarbon and organosiloxane materials.

6.8   Electron Injection Layer

Pure and easily ionized metals such as Ca or Ba can be used as the low-workfunction cathode electron injection layer; however, these metals are highly unstable with gaseous oxygen and water molecules as well as with the organic material used in the electron transport layer. Compounds of such metals that reduce their reactivity and instability while retaining their desired low workfunction include LiF, CsF, Li_2O and Na_2O. Although these are insulators they can be deposited to a thickness of only one or two monolayers and therefore really function to establish the desired interfacial chemistry and a desired trapped interface charge density. This allows the electron injection barrier illustrated in Figure 6.9 to be established.

Other complexes that include alkali metals or alkaline earth metals incorporated in organic molecules are also candidate materials for the electron injection layer. These include *lithium-quinolate* complexes, as shown in Figure 6.17. The development of materials based on organic molecules is of interest due to their inherent compatibility with the other organic layers.

6.9   Hole Transport Layer

The basic requirement of the HTL is good hole conductivity. In conjugated polymers hole conductivity arises through conjugated bonding, and in small-molecule hole transport
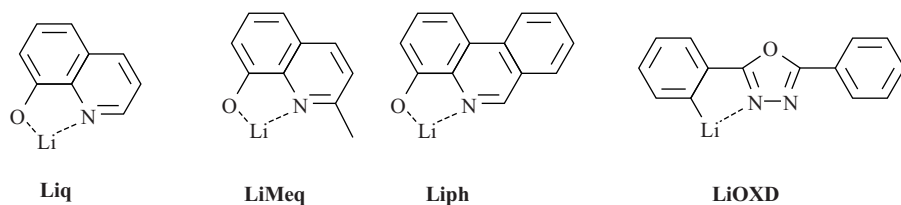
Figure 6.17 *Organo-metallic complexes may also be used for the electron injection layer. Examples are shown consisting of some lithium-quinolate complexes. Liq, LiMeq, Liph and LiOXD. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

materials the same mechanism applies, combined with the transfer of charge between HTL molecules. As shown in Figure 6.11, TPD and NPD are popular hole conductors consisting of small molecules containing six-carbon rings with conjugated bonds allowing intramolecular hole transport. TPD and NPD are members of a family of compounds known as *triarylamenes*.

Triarylamenes were developed for xerography in the 1970s and are well-developed photoconductive materials. Here the electrical conductivity is controlled by the density of mobile charge carriers that are generated by illumination of the triarylamine.

Both TPD and NPD are commonly applied to OLEDs due to their modestly high hole mobilities in the range of 10^{-3} to 10^{-4} cm^2/Vs. A significant challenge is their low-temperature crystallization, which progresses slowly at typical device operation temperatures of 30–40°C. This causes the materials to become mechanically unstable and device stability is compromised.

Another group of triarylamenes include the hole conductors triphenylamine (TPA) and TPTE, shown in Figure 6.18. OLEDs employing these materials may be operated continuously at temperatures of 140°C without breakdown since they do not crystallize readily. A number of other triarylamenes are being studied also. One additional key requirement
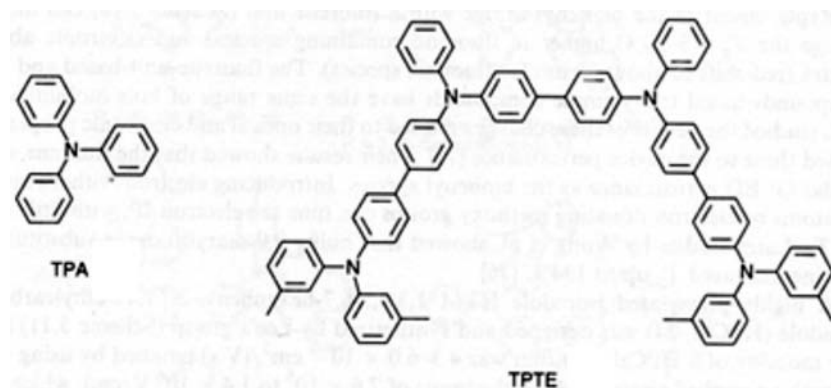
Figure 6.18 *Two further examples of hole-conducting triarylamenes include TPA (triphenylamine) and TPTE (a tetramer of TPA). TPTE enables high-temperature OLED operation without crystallization. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

for efficient OLED devices using them is the size of the energy barrier at the interface of the HTL and the HIL, which must be small enough to result in an efficient OLED. Hence the HOMO level of the HTL should be within a fraction of an electron volt from the anode energy band. In Figure 6.14 the conduction band of ITO is shown at 4.7 eV below a vacuum reference level and the HOMO levels of TPD and NPD are suitable, being close to 5 eV below the vacuum level.

In addition to the triarylamines, another family of hole transport materials consists of the *phenylazomethines*. Four examples of phenylazomethines are shown in Figure 6.19.
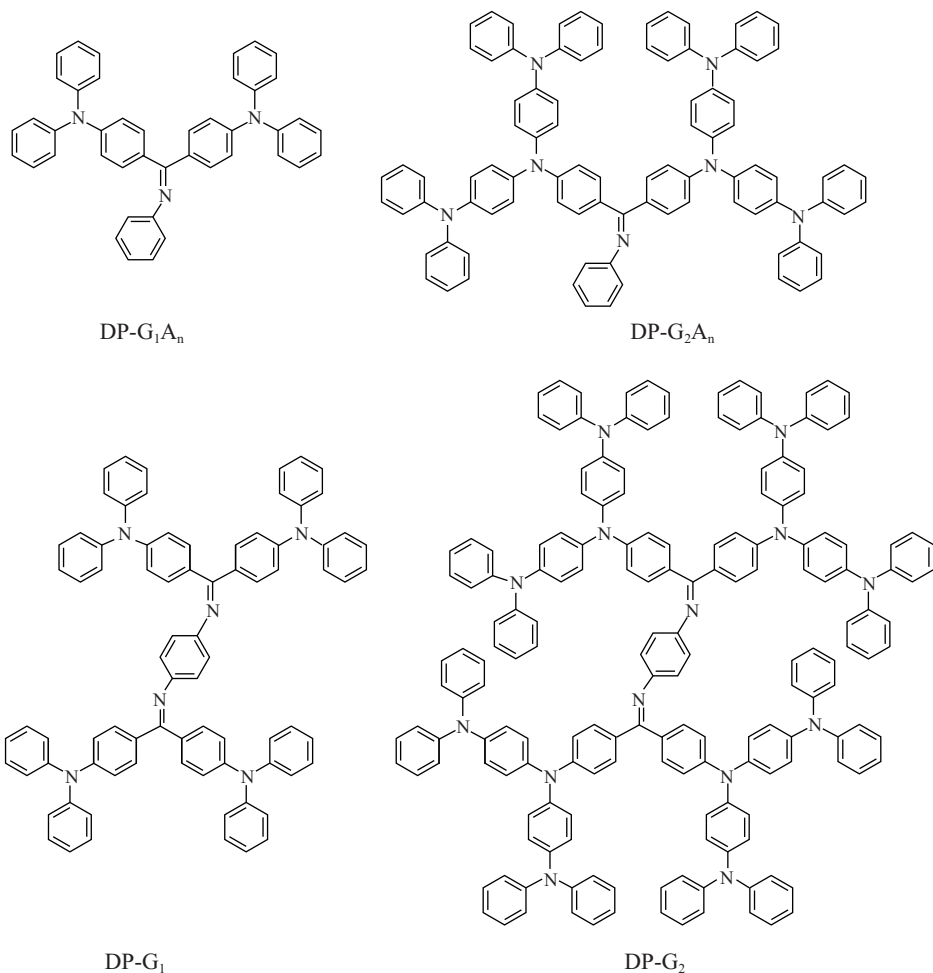
Figure 6.19   *Phenylazomethines are formed by various arrangements of nitrogen-terminated six-carbon rings. These phenylazomethine molecules are thermally stable and are complexed with metal ions such as Sn ions introduced in the form of SnCl$_2$ molecules to form the HTL material. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

Upon mixing these phenylazomethines with metal ions, such as Sn ions, the resulting metal complexes can form good HTL materials with both thermal stability and high-efficiency charge injection of holes. These complexed materials exhibit HOMO levels in the range of -5.2 eV to -5.4 eV (5.2 to 5.4 below the vacuum level), which accounts for their good efficiency.

6.10   Electron Transport Layer

Materials for the electron transport layer (ETL) have been investigated intensively and several families of candidate materials are known. Intermolecular transport occurs by electron hopping, and a LUMO level that is similar in energy to the workfunction of the cathode and the electron-conducting level in the EIL is required, as shown in Figure 6.14.

The ETL should have a mobility of at least 10^{-6} cm^{2}V^{-1}s^{-1}, which is one to two orders of magnitude smaller than the mobility range of HTL materials. Improving this low mobility has been one key target of the intensive investigation of these materials. Insufficient electron mobility in the ETL means that in many cases holes that enter the light emitting layer (LEM) will not encounter electrons and will therefore continue until they reach the ETL before they recombine. Since the ETL is not optimized for high recombination efficiency, a lower device efficiency can result. ETL is also often oxidized by hole conduction, in which electron loss and subsequent degradation of the ETL material occur due to holes that enter the material. This is a major degradation mechanism for small-molecule OLEDs.

The crystallization temperature of the ETL should be high enough to retain the amorphous structure during device operation at the operating temperature. Generally a glass transition temperature should be above 120°C.

The ability of the ETL to withstand long-term exposure to the applied electric field is essential. Since the ETL has a lower mobility and therefore lower conductivity than the HTL a larger voltage drop and hence a larger electric field drops across the ETL. The molecules in the ETL should not lose multiple electrons by field ionization. However, they need to be able to permit the flow of one electron at a time and reversibly change charge state by a single electron charge as the electron enters and leaves a given molecule.

Finally, the ETL material must be able to be processed and coated with good interface stability and with good layer uniformity and quality.

By far the most common and most successful ETL is Alq$_3$ as shown in Figure 6.11. This is an example of a *metal chelate* material. An Al^{3+} ion at the centre of the molecule is surrounded by three side-groups called *quinolines*. Alq$_3$ has a glass transition temperature of over 172°C and an electron mobility of 1.4×10^{-6} cm^{2}V^{-1}s^{-1}. The LUMO level is -3.0 eV below the vacuum level, which is a good match to the cathode workfunction of 3.6–3.8 eV in the case of LiF/Al.

The concern regarding the low mobility of Alq$_3$ is lessened since Alq$_3$ also functions well as a LEM. This means that an oversupply of holes can penetrate a relatively thick ETL and recombine radiatively within this combined LEM/ETL. The emission wavelength in this material has been extensively studied and various substitutions may be made to modify the emission characteristics while retaining the electron transport properties. Examples of substitutions will be discussed in the context of light emitting materials discussed in the next section.
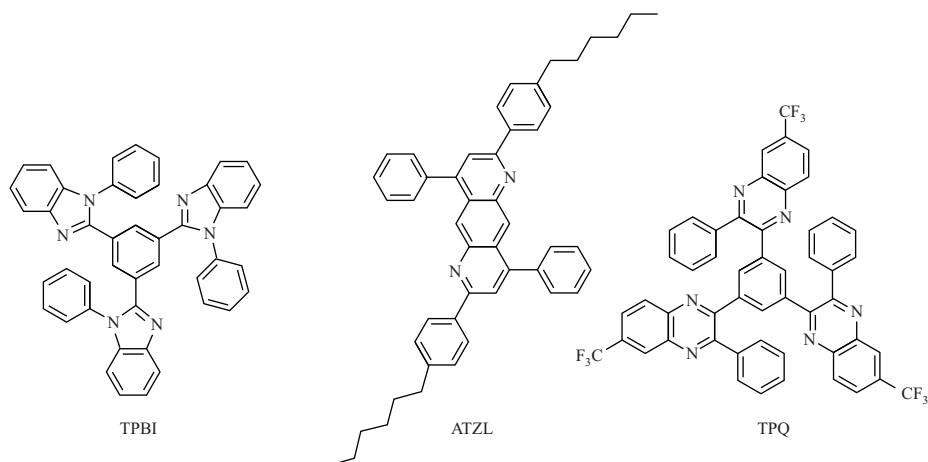
Figure 6.20   *TPBI, ATZL and TPQ are members of imine-based molecules which are candidate electron transport layer (ETL) materials as well as light emission materials. Other candidate ETL materials include C₆₀. See Section 6.16. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

Another important class of ETL materials is the group of *oxadiazoles*. In Figure 6.11 the molecule PBD is an example of an oxadiazole having a LUMO level of $-2.16\,$eV, which permits a high device efficiency. Unfortunately the materials have low glass transition temperatures of about $60°$C, although this can be increased by making larger molecules that resemble groups of two or four PBD molecules connected to each other forming a new molecule with a linear or a star shape respectively. The most serious difficulty associated with the use of oxadiazoles, however, is the tendency of the excited states of the molecules to be unstable resulting in short device lifetimes.

Other potential ETL materials include various molecules containing double-bonded C=N groups, which are known as *imines*. These include TPBI, ATZL and TPQ, as shown in Figure 6.20. TPBI has an electron mobility in the range of 10^{-6} to $10^{-5}\,\text{cm}^2\text{V}^{-1}\text{s}^{-1}$, which is slightly higher than Alq₃. It has a LUMO level of $-2.7\,$eV. TPQ has even higher mobility, of $10^{-4}\,\text{cm}^2\text{V}^{-1}\text{s}^{-1}$ and good thermal stability.

6.11   Light Emitting Material Processes

Obviously a key material for successful OLED operation, the LEM must be amenable to a high-quality deposition technique such as vacuum deposition. It also requires the capability to transport both holes and electrons to enable the recombination of these carriers. Moreover, it must effectively allow for the creation of excitons and their decay to generate photons and it must remain stable at the electric fields needed to transport the holes and electrons and the migration of molecules must be minimized for device stability.

In OLED operation, electrons injected from the cathode and holes injected from the anode combine to form molecular excitons, which were discussed in Chapter 3. The spin of the electron and hole are generally random, which means that the exciton population will occur with a 25% chance of being a singlet exciton and a 75% chance of being a triplet exciton. The singlet exciton can recombine by a dipole emission process and fluorescence will occur; however, the triplet exciton is forbidden to emit dipole radiation. The maximum quantum yield is therefore expected to be 25%.

It is common for mixtures of two or more molecules to be used as light emitting materials in order for the material to provide the various required functions. In such mixtures stability is important and a solid solution of the component molecular materials is usually preferred; the segregation of the components must be avoided. This is commonly referred to as *molecular doping* of one molecular material by another molecular material.

The emission colour of the OLED is ultimately determined by the LEM, and in many cases molecules are modified by changing side-groups, for example, to achieve a specific desired emission colour. In particular, since full-colour displays require efficient red, green and blue light emission, the achievement of LEM emission with suitable colour coordinates for each of the three colours is the goal. In this case a set of three LEMs having the appropriate stability and efficiency and colour coordinate values is required.

A common process used in light emitting materials is the *host-guest* energy transfer process. Here, a host molecule is excited and it can either directly produce radiation or it can transfer its energy to a guest molecule. There are advantages to this energy transfer process which are as follows:

(a) The transfer process can create emission at wavelengths needed for a variety of emission colours without changing the host material.
(b) The 25% theoretical quantum yield can be overcome.

In Section 3.7 energy transfer processes were introduced. The Dexter process involves the transfer of an electron between host and guest molecules and is operational over short-range separations (10–20 Å) only and falls off exponentially. The Förster process is a dipole-dipole interaction process that transfers energy over longer distances of 50 to 100Å. It falls off with the sixth power of the separation R between the two molecules. Finally energy transfer can occur through photon emission from a host molecule followed by photon absorption by a guest molecule. These mechanisms were introduced in Chapter 3.

The host-guest transfer is illustrated in Figure 6.21. S_0^H and S_0^G refer to the singlet ground states of the host and guest respectively. S_1^H and S_1^G refer to excited singlet states of the host and guest respectively. T_1^H and T_1^G refer to the excited triplet states of the host and guest respectively. When a host molecule electron is excited to its S_1^H level it could radiate and emit a photon by a dipole-allowed process and fluoresce. Alternatively the molecule could lose energy by collisions and the electron could return to the ground state and transfer its energy to heat, which is a non-radiative process. In addition the molecule may transfer energy to another molecule by Förster, Dexter or radiative energy transfer, as shown in Figure 6.21. Once transferred, the energy can be radiated from the guest molecule by a dipole-allowed fluorescence process, or it may radiate by phosphorescence. Phosphorescent guest materials are described in Section 6.14. These dopants are particularly interesting since they offer the possibility to improve the quantum yield of the emission process beyond the 25% limit imposed by the 1:3 singlet:triplet ratio.
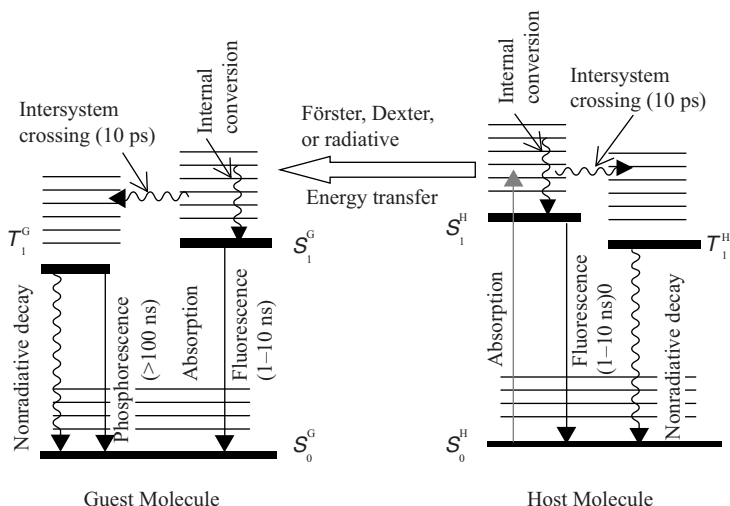
Figure 6.21 *Host-guest energy transfer. The energy transfer can occur due to three possible processes, which may be Förster, Dexter or radiative energy transfer. Energy transfer within either host or guest molecule through intersystem crossing from singlet to triplet excited states can occur by spin-orbit interaction. Reprinted from Li, Z., Meng, H., Organic Light-Emitting Materials and Devices, 157444-574X. Copyright (2006) with permission from Taylor & Francis*

The transfer of energy from host to guest molecules may be a virtually complete transfer or it may involve only a fraction of the excited host molecules. As a result the measured emission spectrum often contains features characteristic of both host and guest emission. This is not desirable when saturated colour coordinates are required for red, green or yellow emission. White OLED emission is desirable for lighting or for monochrome displays, however, and a combined guest-host emission process leads to a broader emission spectrum, which can approximate a desired white spectrum. An important requirement for any energy transfer process is that the host energy is high enough to excite the guest molecule to an excited state. This means that the excited state of the host should be higher in energy than that of the guest.

We can now examine the requirements as well as examples of both host and guest materials.

6.12   Host Materials

Suitable host materials must exhibit good electron and/or hole conduction to ensure the recombination of charge carriers and the effective formation of excitons. Their LUMO and HOMO levels must suit the guest molecules: Compared to the guest LUMO level the host LUMO level should lie less deep in energy (closer to the vacuum level). Compared to the guest HOMO level the host HOMO level should lie deeper (further from the vacuum level) to ensure effective energy transfer. The host and guest molecules must exhibit good miscibility to maintain a stable solution without the tendency for precipitation, which will decrease energy transfer efficiency. Finally energy transfer processes should occur quickly.

Alq$_3$ **63**	BAlq **80**	TPBI **89**	TAZ1 **109**
LUMO: −3.0 eV	LUMO: −3.0 eV	LUMO: −2.7 eV	LUMO: −2.6 eV
HOMO: −5.7 eV	HOMO: −5.9 eV	HOMO: −6.2 eV	HOMO: −6.6 eV
*T1: 2.0 eV	*T1: 2.2 eV (cal)	*T1: XXX	*T1: XXX
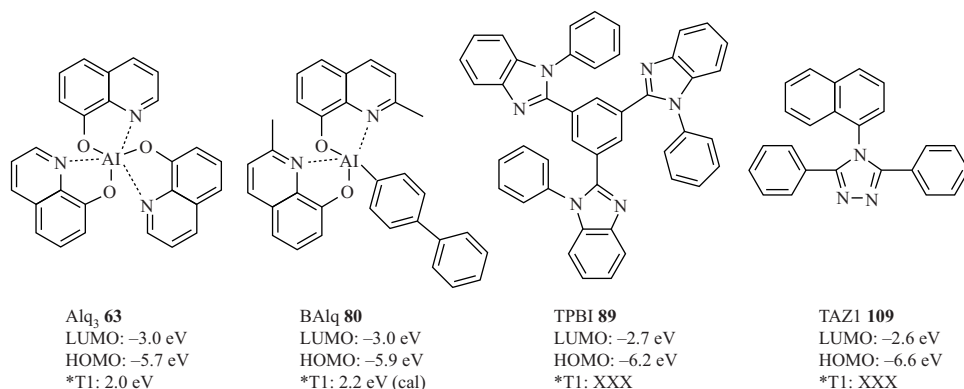
Figure 6.22 *Electron transport hosts Alq$_3$, BAlq, TPBI and TAZ1. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

Among the simplest LEM hosts is the single molecule Alq$_3$, which also functions as an ETL material. Alq$_3$ emits in a band centred near 560 nm, which results a yellow-green emission colour. Since the emission spectrum is not an appropriate red, green or blue colour for a full-colour display it can be combined with other emission spectra to yield white emission for white-emitting OLED lamps. It can also be optically filtered to achieve red or green emission. Filtering does reduce efficiency, however, as unwanted parts of the emission spectrum are absorbed by the filter. The LUMO and HOMO levels of Alq$_3$ are −3 eV and −5.7 eV respectively, and its triplet energy is 2 eV. One modification of Alq$_3$ is BAlq, which has a LUMO level of −3 eV and a HOMO level of −5.9 eV. BAlq has a slightly higher triplet energy of approximately 2.2 eV compared to Alq$_3$, which makes it an ideal host for some red-emitting phosphorescent guest molecules, which will be described in the next section. The relevant LUMO and HOMO energy levels of Alq$_3$, BAlq as well as two other hosts that favour electron transport are shown in Figure 6.22.

In order to provide hole transport within the LEM, hole transport hosts are often combined with these electron transport hosts. A number of effective hole transport hosts are known. They include CPB, with HOMO and LUMO levels of −3 eV and −6 eV respectively, and a triplet level of 2.67 eV. CPB works well for red, yellow and green triplet emitters, but the triplet level is not high enough for blue emission. For blue emission, CDBP has been shown to work well, with HOMO and LUMO levels of −3 eV and −6.3 eV, and triplet level of 3.0 eV. Both CPB and CDBP are shown in Figure 6.23.

6.13   Fluorescent Dopants

The requirements for full-colour display applications of OLEDs include red, green and blue emitters with colour coordinates close to the following values: for green emitters, $x = 0.3$ and $y = 0.6$; for red $x = 0.62$ and $y = 0.37$; for blue $x = 0.14$ and $y = 0.10$. Fluorescent dopants emitting with approximately these colour coordinates are required.

CBP **110**
LUMO: −3.0 eV
HOMO: −6.3 eV
*T1: 2.67 eV

CDBP **111**
LUMO: −3.0 eV
HOMO: −6.3 eV
*T1: 3.0 eV
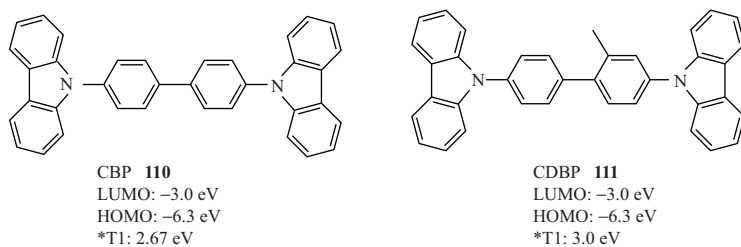
Figure 6.23   Hole transport hosts CBP and CDBP. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis

An example of a green dopant is based on a *coumarin* dye molecule such as C-545TB, shown in Figure 6.24. This dopant yields saturated green emission with colour coordinates $x = 0.3$, $y = 0.64$, with a luminescent efficiency of 12.9 cd A^{-1}, a power efficiency of 3.5 lm W^{-1} at 20 mA cm^{-2} and a brightness of 2585 cd m^{-2}. Another type of green dopant is DMQA, which is an example of a *quinacridone* molecule. DMQA achieved a luminescent efficiency of 21.1 cd A^{-1} and a luminance of over $88\,000$ cd m^{-2}. Coumarin-based C-545TB and quinacridone-based DMQA are shown in Figure 6.23.

Red fluorescent dopants have been developed that simultaneously exhibit satisfactory colour coordinates with good stability and efficiency based on the *arylidene* family of molecules. An example of a red fluorescent molecule is DCJPP, shown in Figure 6.25. There are large numbers of other candidate red fluorescent materials in this family; however, the less suitable ones suffer from a tendency to undergo unwanted chemical reactions. Some fluoresce with *y*-values of their colour coordinates that are too large and orange-red emission results. Still others exhibit deep-red emission but have low quantum efficiencies.

Red fluorescent molecules based on other molecular families exist. For example, an *isophorone*-based red emitter, DCDDC (see Figure 6.25), has been used as a red emitter in OLEDs when dissolved in the host Alq$_3$. Emission from both Alq$_3$ and DCDDC is observed for small concentrations of dopant; however, if the DCDDC doping level is increased to above 2% concentration, only the red emission is observed due to a strong

C-545TB (R$_1$ = CH$_3$, R = t-butyl; R′ = H)
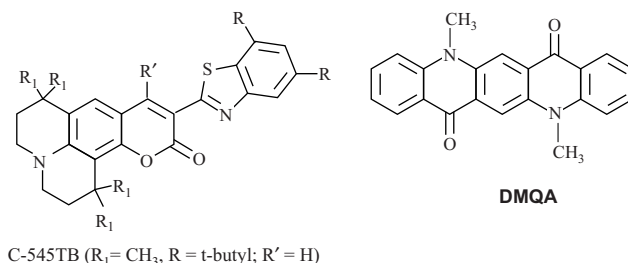
DMQA

Figure 6.24   Coumarin-based green fluorescent dopant C-545TB and quinacridone-based dopant DMQA. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis
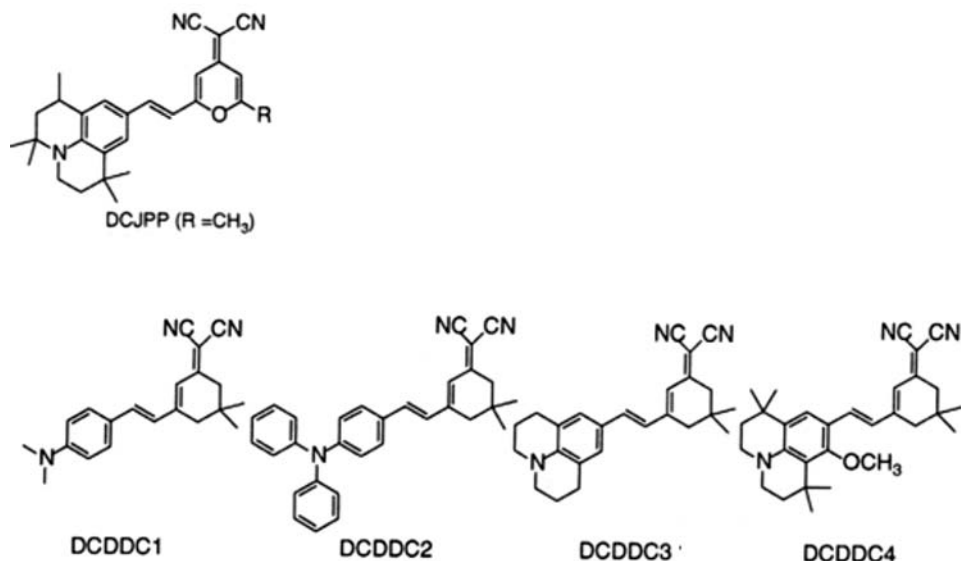
Figure 6.25 *The red fluorescent molecule DCJPP derived from the arylidene family of molecules and four variations of red fluorescent molecule DCDDC derived from the isophorone family of molecules. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

host-to-guest energy transfer. The emission peak is 630 nm from the DCDDC. At a 1% DCDDC concentration, which does somewhat compromise the red colour, a peak luminance of 5600 cd m^{-2} at a voltage of 15 V with maximum efficiency of 1.6 lm W^{-1} is achieved.

There has been a great deal of effort invested in blue fluorescent molecules and suitable hosts. The challenge is that the short wavelength of emission in the range of 450 nm required for blue colour coordinates with y-values near 0.1 calls for high bandgaps between guest LUMO and HOMO levels and even higher bandgaps near 3 eV between suitable host LUMO and HOMO levels. Molecules with these properties exist but the resulting OLEDs have proven less stable than red and green emitters.

An example of a blue emitter host that is a *distyrylaraline* derivative is DPVBI, having HOMO and LUMO levels of -5.9 eV and -2.8 eV respectively, and bandgap of 3.1 eV. A suitable fluorescent guest that is also a distyrylaraline derivative is BCzVBI, with HOMO and LUMO levels of -5.4 eV and -2.42 eV. Luminance centred at 468 nm at 10 000 cd m^{-2} at an efficiency of 0.7–0.8 lm W^{-1} has been achieved (see Figure 6.26).

Also shown in Figure 6.26 is a candidate blue system from the *anthracene* family. The host is JBEM with HOMO and LUMO levels -5.8 eV and -2.8 eV respectively, and the guest is the well-known *perylene* molecule with HOMO and LUMO levels -5.3 eV and -2.5 eV respectively. Resulting OLED performance achieves 400 cd m^{-2} at a current density of 20 mA cm^{-2}, a maximum efficiency of 1.45 lm W^{-1} and colour coordinates of $x = 0.24$ and $y = 0.21$. A half-life (life to half initial luminance) starting at 100 cd m^{-2} brightness of over 1000 hours can be obtained. Improvements to lower the y-component

DPVBI

BCzVBI

JBEM
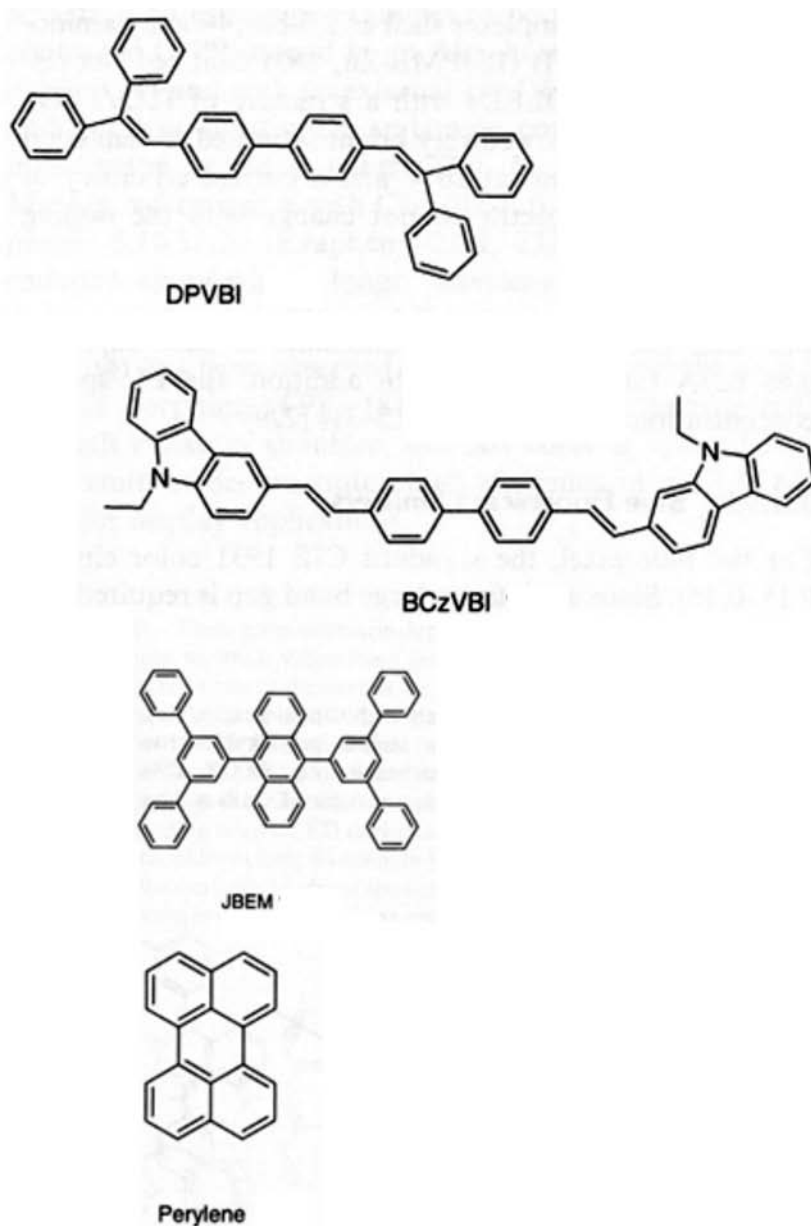
Perylene

Figure 6.26 *Distytylaraline family host DPVBI and dopant BCzVBI. Also shown are anthracene family host JBEM and dopant perylene. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

of the colour coordinate and achieve pure blue emission can be realized by using other anthracene derivatives.

6.14 Phosphorescent Dopants

Among the highest efficiency dopants are the iridium organometallic complexes. They have a short triplet lifetime of 1–100 μs, which means that radiative recombination is assisted since the normally forbidden radiation from the triplet exciton is somewhat allowed due to spin-orbital interaction in the molecule. This relaxes the requirement that spin is invariant during the transition and triplet excitons become allowed radiative transitions. High-efficiency phosphorescence results. Iridium, a transition metal with an unfilled inner shell and a net angular momentum, provides the needed spin-orbit interaction.

Examples of red, green and blue iridium-based emitters are shown in Figure 6.27, although numerous phosphorescent emitters are well studied and some have been commercialized in OLEDs for battery-powered devices that require high efficiency.

6.15 Organic Solar Cells

The absorption of sunlight in a molecular organic semiconductor results in the formation of molecular excitons, and in accordance with the dipole process and the discussion in Sections 3.6 and 3.7 the ground-state singlet is excited into an excited singlet molecular exciton. This exciton is localized to a single molecule, which is generally on the nanometer length scale. Unless the exciton can be dissociated and its hole and electron extracted no current can result. In contrast to this, photon absorption in inorganic semiconductors used in solar cells results in separated holes and electrons that are free to flow independently of each other and hence directly contribute to current flow.

A key challenge in the development of organic solar cells is to overcome the localization and pairing in the form of excitons of optically generated holes and electrons. Once dissociated, charges can flow from molecule to molecule by a hopping process. Materials and device architectures designed to facilitate exciton dissociation are the key to successful organic solar cells.

The simplest organic solar cell structure is the single-layer device shown in Figure 6.28. Photons create molecular excitons in the organic semiconductor layer. The device relies on the differing workfunctions between cathode and anode to generate an electric field high enough to collect these charges. The diffusion length of carriers is generally on the order of 10 nm and the thickness of the semiconductor layer is much more than 10 nm to achieve a reasonable degree of photon absorption. The majority of the generated excitons are never dissociated and only a small fraction of generated carriers are collected by the electrodes. In fact the electric field is only high right at the abrupt electrode-semiconductor interfaces, and this is where exciton dissociation can most readily occur. Efficient carrier collection, however, requires the participation of both carrier types, which is not favoured by this approach. The energy level diagram for this structure is shown in Figure 6.29.
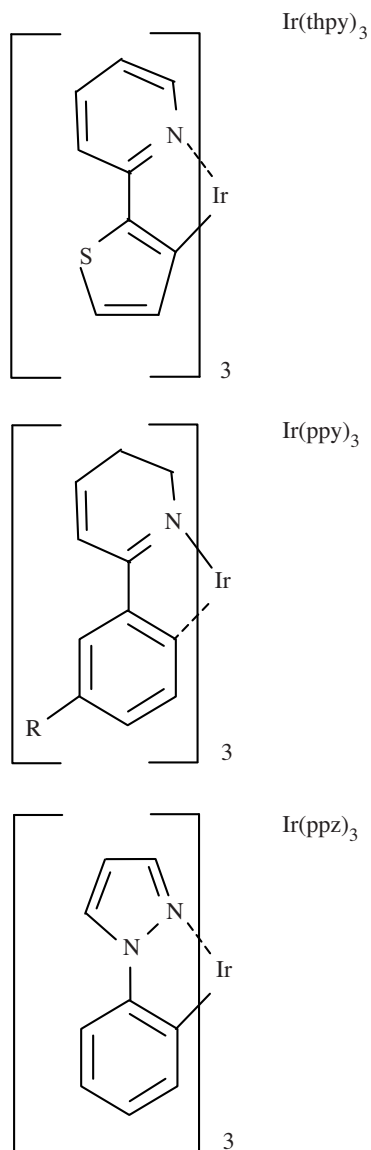
Figure 6.27 *Phosphorescent iridium-based emitters: Red- Ir(thpy)$_3$, Green-Ir(ppy)$_3$, Blue-Ir(ppz)$_3$. Chemical Structure reproduced from Organic light-emitting materials and devices, ed by Z. Li and H. Meng 9781574445749 (2007) Taylor and Francis*

Cathode
Organic semiconductor
ITO
Glass

Figure 6.28 *Single-layer organic solar cell consisting of a single organic semiconductor layer, a low workfunction cathode and a transparent anode. Device efficiency is well below 1%*

A substantial increase in the collection of charge may be accomplished by the *planar heterojunction* solar cell, which is shown in Figure 6.30. Its energy band diagram is shown in Figure 6.31. The introduction of an electron transport *acceptor* layer (ETL) and a hole transport *donor* layer (HTL) creates a strong electric field at the heterojunction interface that greatly enhances exciton dissociation there. In principle both the donor and acceptor layers can absorb photons and become populated with excitons. These excitons can then diffuse towards the heterojunction interface and dissociate there.

The sharp and narrow high field region at the heterojunction interface in organic solar cells may be contrasted with the inorganic p-n junction. The width of the depletion region in inorganic junctions is determined by the spatial extent over which carriers recombine to establish an equilibrium condition. At the organic interface in organic junctions, charge carriers in the HOMO and LUMO levels transfer from molecule to molecule by hopping and only minimal charge transfer occurs, leaving the equivalent of a depletion region of very small thickness. The potential difference between the HOMO and LUMO levels falls across a very small spatial range of dimension in the nanometer scale giving rise to a high electric field at the junction.

In practice, however, the donor layer is specifically designed to absorb photons and become populated with excitons. Since the mobility of holes is relatively higher than electrons, as we saw in Section 6.10, the donor HTL allows the holes to diffuse towards the heterojunction interface. The electrons will remain bound to the holes since exciton
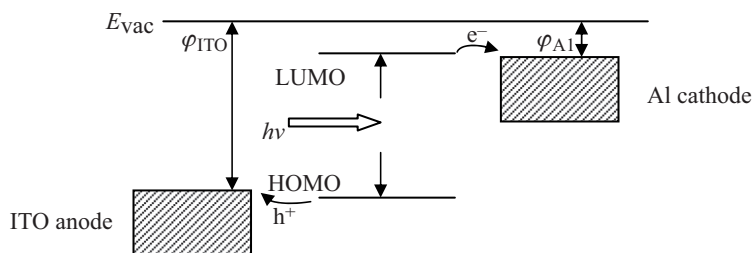
Figure 6.29 *Energy level diagram for single-layer organic solar cell. The absorption of light creates excitons through the promotion of molecular electrons from the HOMO level to the LUMO level. Electrode workfunctions are set to match the HOMO and LUMO levels to facilitate the collection of the electrons and holes as shown*

Cathode
Organic acceptor ETL material
Organic donor HTL material
ITO
Glass

Figure 6.30   *Organic planar heterojunction solar cell structure showing donor and acceptor organic layers*

dissociation will not readily take place until the excitons reach the interface. This interface now enables the collection of both electrons and holes, which drift across their respective layers: Holes reach the ITO electrode through the hole-conducting layer, and electrons reach the cathode through the electron-conducting layer.

The terminology 'donor' and 'acceptor' used to describe the two layers forming the heterojunction comes about since electrons that are dissociated from the excitons in the donor layer at the junction are transferred or donated across this junction from donor molecules and accepted by acceptor molecules in the acceptor layer. The holes from the dissociated excitons remain in the donor layer and drift to the anode. The terminology is a molecular analogue of the terms 'donor' and 'acceptor' applied to dopants used in inorganic semiconductors; however, the organic molecules donate and accept electrons to/from neighbouring molecules rather than to/from energy bands. Since the acceptor layer becomes populated with electrons this layer needs to be an electron conductor and charge
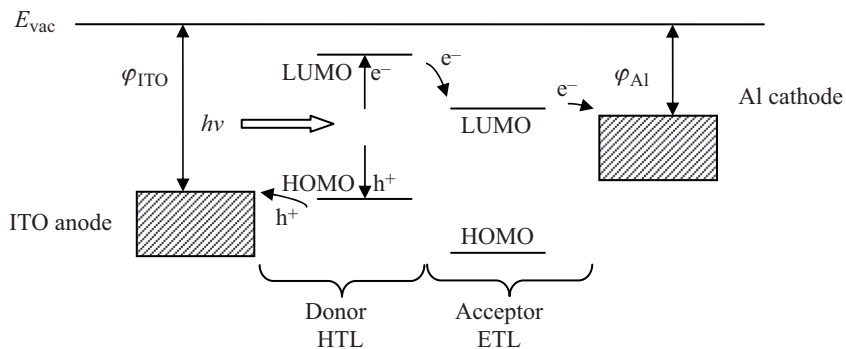
Figure 6.31   *Heterojunction solar cell showing donor and acceptor LUMO and HOMO levels. Excitons are generated throughout the donor layer and these excitons are dissociated when they diffuse to the donor-acceptor interface. Finally the separated holes and electrons can drift to their respective electrodes*

is carried in its LUMO level. Conversely the donor layer, being populated with holes, needs to be a hole conductor and these holes are carried in its HOMO level.

The thickness of the p-type donor layer is controlled by the diffusion length of the excitons which must reach the interface to be dissociated. A donor layer that is too thick will lower efficiency since a significant fraction of the generated excitons will recombine before they can reach the interface. A layer that is too thin will result in less absorption of light. A solar cell efficiency of only a few percent is achievable with the planar heterojunction design.

Since diffusion lengths in organic materials are approximately 10 nm, the useful absorption depth in the donor layer is only about 10 nm, which means that incomplete absorption of sunlight limits the performance of the heterojunction solar cell. The thickness for virtually complete absorption of sunlight is closer to 100 nm in organic materials; however, a donor layer of this thickness would result in poor efficiency and most generated excitons would recombine without reaching the interface.

A successful approach to improving performance further is to arrange several interfaces within the light path of the incoming sunlight, and to make each donor layer thin enough to allow effective exciton diffusion to the nearest heterojunction interface. A portion of sunlight is absorbed in each thin donor layer and the remaining light can then continue to a subsequent layer. This approach relies on a *bulk heterojunction* layer that incorporates multiple donor and acceptor regions. The device structure is as shown in Figure 6.32.

The bulk heterojunction layer can be formed using a variety of nanostructures, and the development of techniques and materials for the achievement of these nanostructures has been a focal point in further improving organic solar cells. Bulk heterojunction solar cells have attained the highest efficiency levels available in organic solar cells that use a single organic absorption band to absorb light.

The length scale of the desired nanostructures is in the nanometer range, and it is very desirable to use *self-organization* of the organic materials to achieve a low-cost method to create these nanostructures. For example, the donor and acceptor organic materials can be mixed together and then deposited onto the solar cell substrate. If the mixed material segregates spontaneously under suitable conditions to form the desired bulk heterojunction then self-organization has been achieved. This dramatically lowers the cost of processing since submicron lithography and patterning techniques are avoided. Although these techniques are well known and highly developed for inorganic semiconductor device processing they are not cost effective for large-area solar cells.

| Cathode |
| Bulk heterojunction |
| ITO |
| Glass |

Figure 6.32 *Bulk heterojunction organic solar cell. A number of small (∼10 nm) donor regions are organized within the bulk heterojunction layer and optimized to absorb sunlight and allow exciton diffusion to a nearby junction*

An important requirement of bulk heterojunctions is to provide for the effective conduction of current away from the donor and acceptor regions and for the collection of this current by the electrodes. Two specific examples of heterostructures are shown in Figure 6.33. Figure 6.33a shows a morphology that limits the effectiveness of current collection because the donor and acceptor layers are not well connected to the electrodes. This is not an accurate representation of the morphology in real systems, but serves for illustration purposes. Figure 6.33b shows a more desirable structure since the donor and acceptor materials are arranged to allow for effective connection to the electrodes. The achievement of organic layers that self-organize into optimal structures at low cost is an area of ongoing research.
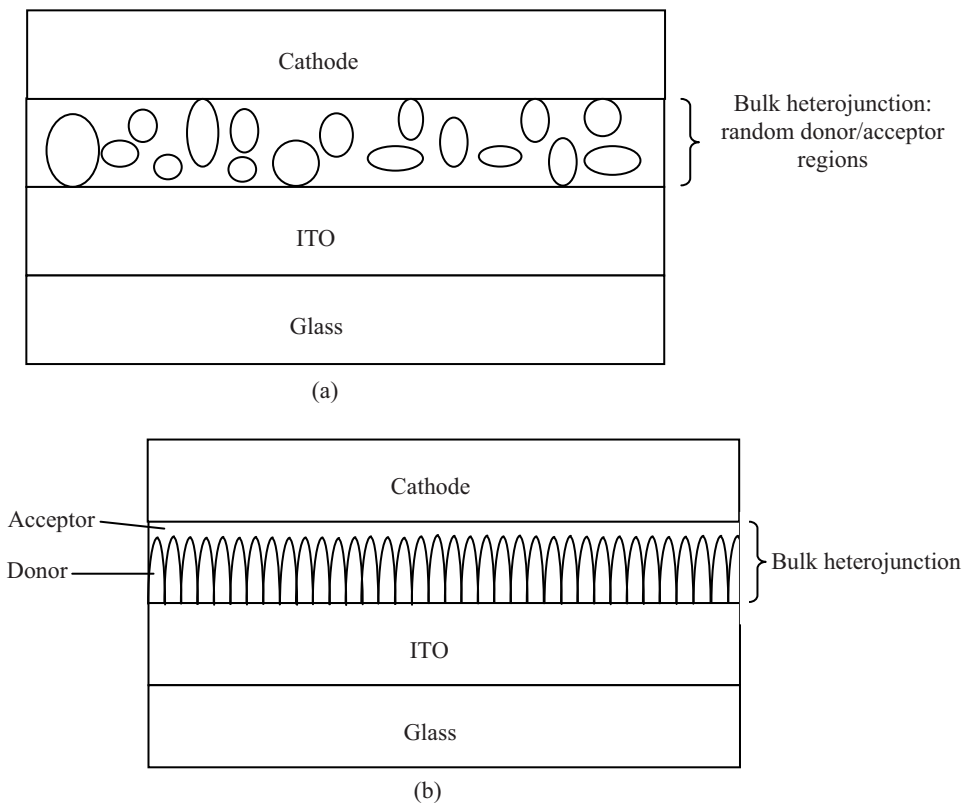
Figure 6.33 *(a) Bulk heterojunction structure showing a typical random structure of donor and acceptor materials. The dimension of one region within the heterojunction is about 10 nm. The problem is the connectivity of these regions to their appropriate contact materials. (b) Bulk heterojunction of vertically oriented stripes of donor and acceptor materials that enables the donor material to be in contact with the ITO electrode and the acceptor layer to be in contact with the aluminium electrode. The acceptor layer could be made using vertically oriented carbon nanotubes*

Another important area of current development is the use of multilayer organic solar cells in which various layers act to absorb different portions of the solar spectrum in a manner analogous to the inorganic multiple junction solar cells described in Sections 4.11 and 4.13. This is particularly important for organic solar cells since the absorption bandwidth of a given organic material is small. This is because the π and π^* bands are much narrower than the conduction and valence bands in inorganic semiconductors, which limits the absorption bandwidth.

The heterojunction structures we have discussed are capable of providing two absorption bands if excitons can be generated and harvested in both the donor and acceptor materials. The energy gaps of these two layers can be different and two absorption bands can be realized. A challenge associated with this is achieving high enough diffusion lengths in both the donor and acceptor layers and collecting carriers effectively. A single organic material having all the attributes needed specifically for an ideal acceptor, including good electron mobility, high optical absorption, and effective electron capture from the donor material, has not yet been found. Nevertheless numerous organic material blends and mixtures are being investigated to obtain multiple absorption bands in organic solar cell structures.

To date, efficiencies as high as 7–8% have been reached in the laboratory for organic solar cells. The attributes of low cost, low weight and flexibility are the key drivers behind this development, although the available materials ultimately control the achievement of high performance. A few popular organic materials will now be reviewed.

6.16   Organic Solar Cell Materials

A common *thiophene*-based donor material is poly(3-hexylthiophene), or P3HT. It is soluble in several organic solvents, which makes it compatible with low-cost solution processing. Another donor material is poly(3,3‴-didodecyl quaterthiophene), or PQT-12. Their molecular structures are shown in Figure 6.34. The optical absorption spectra of these compounds (Figure 6.35) show that their absorption bands are limited to the green and red parts of the solar spectrum.

Acceptor materials must provide good electron conductivity but optical absorption is not desired. A popular material is a C_{60} derivative. C_{60} is a fullerene, or a molecule composed entirely of carbon. The graphite-like surface of C_{60} allows effective electron transport through the delocalized electrons in the molecular orbitals, which result in the LUMO level. C_{60} as well as its derivative [6,6]-phenyl-C_{61}-butyric acid methyl ester (PCBM) are shown in Figure 6.36. The modification allows the fullerene derivative PCBM to be solution-processed, which reduces manufacturing costs, whereas C_{60} must be vacuum deposited.

Fullerenes are excellent acceptors since they have a LUMO level with electron energy well below the vacuum level and more specifically somewhat below the LUMO levels of a variety of donor molecules as required for solar cells. This is a requirement of an acceptor molecule, as illustrated in Figure 6.31. Each C_{60} actually has the ability to accept multiple electrons due to the number of available vacancies in the LUMO level. Electron transport within a single C_{60} molecule is very fast and efficient. The electrons are metastable within

(a)   (b)
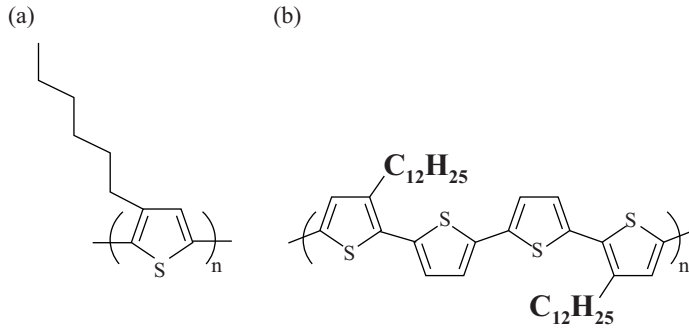
Figure 6.34 *Molecular structures of poly(3-hexylthiophene) (or P3HT) and PQT-12. Reprinted with permission from Organic Electronics, Efficient bulk heterojunction solar cells from regio-regular-poly(3,3'''-didodecyl quaterthiophene)/PC70BM blends by P. Vemulamada, G. Hao, T. Kietzke and A. Sellinger, 9, 5, 661-666 Copyright (2008) Elsevier*

the molecules and are therefore readily transferred to the solar cell electrode. Finally these fullerenes have an absorption spectrum that peaks in the ultraviolet part of the spectrum, which means that they exhibit only minimal absorption in the important parts of the solar spectrum, leaving the donor layer free to absorb the visible or infrared solar radiation.

The *carbon nanotube* is another acceptor material of considerable interest. A promising development is the use of carbon nanotubes as rod-like acceptors to form structures similar to that shown in Figure 6.33b. Nanotubes, being composed of rolled-up graphite sheets, have electronic properties that are similar to the fullerenes, allowing them to function
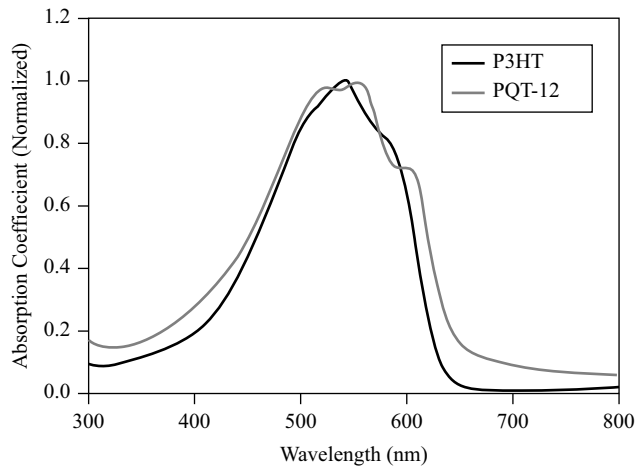
Figure 6.35 *Absorption spectra of P3HT and PQT-12. Reprinted with permission from Organic Electronics, Efficient bulk heterojunction solar cells from regio-regular-poly(3,3'''-didodecyl quaterthiophene)/PC70BM blends by P. Vemulamada, G. Hao, T. Kietzke and A. Sellinger, 9, 5, 661-666 Copyright (2008) Elsevier*
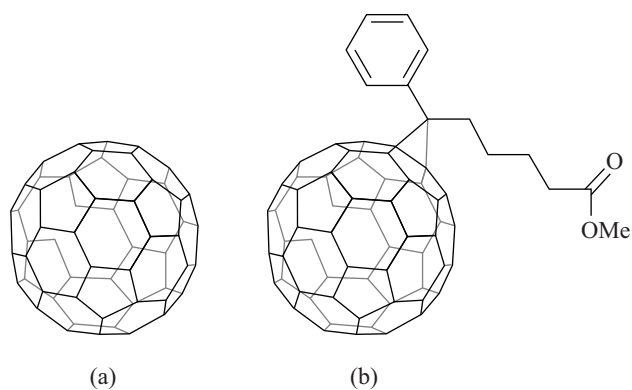
Figure 6.36 *Molecular structures of the fullerene C$_{60}$ and its derivative [6,6]-phenyl-C$_{61}$-butyric acid methyl ester (PCBM). Reproduced from James Hugh Gervase Owen, Nanotubes and fullerenes for Quantum Computing, http://homepage.mac.com/jhgowen/research/nanotube_page/nanotubes.html. Copyright (2011) with permission from J. H. G. Owen*

effectively as acceptors. The achievement of bulk heterojunctions with oriented nanotubes is required. A carbon nanotube is shown in Figure 6.37.

There are numerous new materials being studied and the list of these is growing steadily. The field of organic solar cells is in its infancy and research is underway in materials and nanostructures to optimize optical absorption, carrier collection efficiency, operational stability and large-scale manufacturability. The measured efficiency of organic solar cells is rising rapidly (see Figure 4.6). Very recent results not shown in Figure 4.6 have demonstrated over 8% efficiency for organic solar cells using newly developed materials.
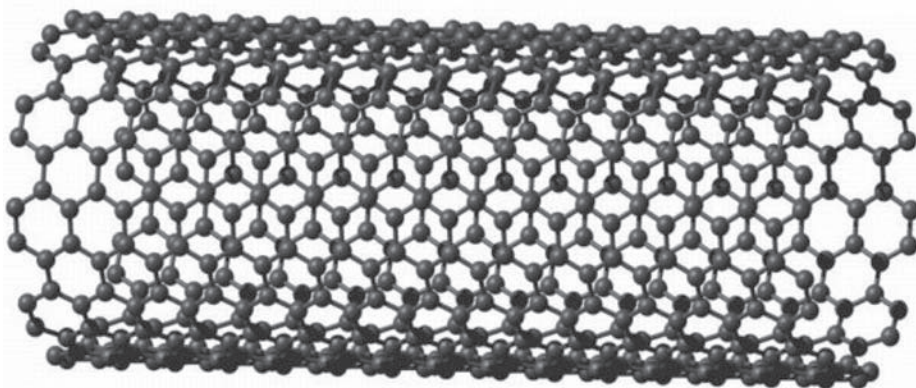
Figure 6.37 *Carbon nanotube. Courtesy of Dr. J.H.G Owen and the Oxford University QIP-IRC*

6.17   Summary

6.1. The field of organic electronics is progressing rapidly and is primarily driven by the traditional advantages of polymers in terms of flexibility and lightness with low manufacturing costs. It is based on specific polymers with electronic properties that were identified in the 1970s.

6.2. Electrical conductivity in molecular materials is made possible by conjugated bonding in which electrons are delocalized in π sub-bands. Electron transport within molecules occurs by intramolecular transport, and transport between molecules occurs by intermolecular hopping. Conjugated polymers allow for numerous modifications that can modify bandgaps as well as side-groups that facilitate solution processing.

6.3. In its simplest form the polymer OLED uses one electroluminescent polymer layer that acts as an electron transport layer near the cathode and as a hole transport layer near the anode. The EL polymer layer also provides for the recombination of electrons and holes that form molecular excitons within this layer. In an energy diagram an electric field forms between anode and cathode, which is the result of a charge transfer. Electrons from the cathode drift by means of the lowest unoccupied molecular orbital, or LUMO, while holes drift from the anode by means of the highest occupied molecular orbital, or HOMO. Molecular excitons and electroluminescence result.

6.4. The small-molecule OLED uses separate electron- and hole-transporting materials. A hole transport layer (HTL) and a separate electron transport layer (ETL) form a junction sandwiched between anode and cathode electrodes. The cathode electrode has a low workfunction to allow electrons to readily flow into the ETL. A light emitting material (LEM) positioned between the HTL and ETL is optimized for the formation of molecular excitons, and their radiative recombination efficiency is also optimized. Other layers may be added to further optimize device operation.

6.5. Anode materials are normally transparent to allow light emission, and indium tin oxide (ITO) is the most common material. Surface smoothness of both the substrate and the ITO layer is important. A surface roughness below 2 nm is generally required although ITO is inevitably rough due to its polycrystalline structure. ITO has a high workfunction ($\varphi > 4.1$ eV) allowing it to inject holes efficiently. It has a chemically active surface that can cause migration of indium into subsequent polymer layers.

6.6. Cathode materials must provide high conductivity, a low workfunction and good adhesion to the underlying polymer layers with long-term stability. Ease of oxidation and a tendency for these cathode layers to cause chemical reduction of adjacent polymer layers are challenges. A two-layer cathode is popular and the LiF/Al structure is widely used in which the aluminium protects the reactive LiF layer and also provides improved sheet conductivity. OLED devices must be protected with encapsulation. The rate of moisture penetration must be calculated to ensure a specified product life.

6.7. The hole injection layer (HIL) acts to improve the smoothness of the anode surface due to the native ITO roughness and facilitates efficient hole injection and long-term hole injection stability. The polymer-metal interface is very complex due to charges that are trapped there as a result of dangling bonds and metal atoms that react with the

organic layer. Another important aspect of OLED performance relates to the current balance between electron and hole currents.

6.8. Pure and easily ionized metals such as Ca or Ba can be used as the low workfunction cathode electron injection layer; however, these metals are highly unstable with gaseous oxygen and water molecules as well as with the organic material used in the electron transport layer. Compounds of such metals that reduce their reactivity and instability while retaining their desired low workfunction include LiF. Although these are insulators they can be deposited to a thickness of only one or two monolayers. The development of materials based on organic molecules is of interest due to their inherent compatibility with the other organic layers.

6.9. The hole transport layer (HTL) must provide effective hole transport. TPD and NPD are popular hole conductors consisting of small molecules containing six-carbon rings with conjugated bonds allowing intramolecular hole transport. TPD and NPD are members of a family of compounds known as triarylamenes and have hole mobilities in the range of 10^{-3} to 10^{-4} cm^2 V^{-1} s^{-1}. A challenge is their low-temperature crystallization. OLEDs employing TPA and TPTE may be operated continuously at temperatures of 140°C without breakdown since they do not crystallize readily. A number of other materials are being studied also.

6.10. The electron transport layer (ETL) provides electron transport. A LUMO level that is similar in energy to the workfunction of the cathode is required. The ETL should have a mobility of at least 10^{-6} cm^2 V^{-1} s^{-1}, which is one to two orders of magnitude smaller than the mobility range of HTL materials. Improving this low mobility has been a key target. Since the ETL has a lower mobility it must support a larger electric field, which can further compromise stability. The most common and most successful ETL is Alq$_3$.

6.11. Several processes occur to achieve light emission in the light emitting material (LEM). It must be able to transport both holes and electrons to enable the recombination of these carriers. It must effectively allow for the creation of excitons and their decay to generate photons. It must also remain stable at the electric fields needed to transport the holes and electrons, and the migration of molecules must be minimized for device stability. It is common for mixtures of two or more molecules to be used as light emitting materials in order for the material to provide the various required functions, including the host-guest energy transfer process in which an excited host molecule can either directly produce radiation or transfer its energy to a guest molecule. This can overcome the 25% efficiency limit of singlet recombination.

6.12. Suitable host materials must exhibit good electron and/or hole conduction and their LUMO and HOMO levels must match the guest molecules. They must exhibit good miscibility to maintain a stable solution without the tendency for precipitation, which will decrease energy transfer efficiency. Finally, energy transfer processes to guest molecules should occur rapidly. Alq$_3$ is a simple host material that also functions as an ETL material. It emits at 560 nm, which is a yellow-green colour, and it can be combined with other guest emission spectra to yield a white emission or it can be optically filtered to achieve red or green emission.

6.13. Guest materials may fluoresce or photoluminesce. Fluorescent dopants emitting with suitable red, green and blue colour coordinates are required for full-colour displays. In

addition the HOMO and LUMO levels of the dopants must match the host materials. Guest phosphorescent dopants should have triplet energy levels smaller than the host to enable effective energy transfer. The guest dopant should remain soluble in the host to prevent unwanted segregation. An example of a green dopant is based on a coumarin dye molecule. Red fluorescent dopants have been developed that simultaneously exhibit satisfactory colour coordinates with good stability and efficiency based on the arylidene family of molecules. For blue emission, hosts from the anthracenes combined with the guest molecule perylene may be used.

6.14. For phosphorescent dopants iridium organometallic complexes have proven the most effective molecules. They have a short triplet lifetime of $1–100 \,\mu s$. Radiative recombination is assisted since the normally forbidden radiation from the triplet exciton is somewhat allowed due to spin-orbital interaction in the molecule. This alters the spin states and renders high-efficiency phosphorescence. Numerous phosphorescent emitters are well studied and some have been commercialized in OLEDs for battery-powered devices that require high efficiency.

6.15. In the organic solar cell, absorption of sunlight creates molecular excitons that are localized to a single molecule, which is generally on the nanometer length scale. Unless the exciton can be dissociated and its hole and electron extracted, no current can result. A key challenge is to overcome the localization and pairing in the form of excitons of optically generated holes and electrons. Materials and device architectures designed to facilitate exciton dissociation are the key to successful organic solar cells. The single-layer solar cell relies on the differing workfunctions between cathode and anode to generate an electric field high enough to collect these charges. Better performance may be achieved in heterojunction and bulk heterojunction devices in which excitons are dissociated at heterojunctions between donor and acceptor materials.

6.16. The thiophene P3HT and fullerene C_{60} are simple examples of a donor material and acceptor material respectively. Acceptor materials must provide good electron conductivity, but optical absorption is not desired. Electron transport within a single C_{60} molecule is very fast and efficient. The electrons are metastable within the molecules and are therefore readily transferred to the solar cell electrode. Finally, these fullerenes have an absorption spectrum that peaks in the ultraviolet part of the spectrum, which means that they exhibit only minimal absorption in the important parts of the solar spectrum, leaving the donor layer free to absorb the visible or infrared solar radiation. The carbon nanotube is another acceptor material of considerable interest.

Suggestions for Further Reading

So F. *Organic Electronics – Materials, Processing, Devices, and Applications*. CRC Press, 2010.

Li Z and Meng H. *Organic Light Emitting Materials and Devices*. CRC Press, 2007.

Sun SS and Dalton LR. *Introduction to Organic Electronic and Optoelectronic Materials and Devices*. CRC Press, 2008.

Problems

6.1 The well-known firefly produces light by a process called bioluminescence. Find more information on the specific organic molecules in the firefly and the way they are excited.

6.2 The benzene series is shown below:
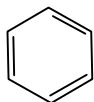
Benzene, absorption about 255 nm
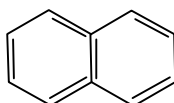
Naphthalene, absorption about 315 nm
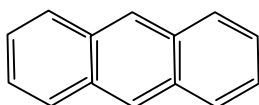
Anthracene, absorption about 380 nm
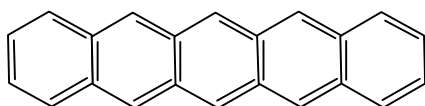
Pentacene, absorption about 580 nm

(a) Look up and reproduce absorption spectra for the four members of the benzene series shown.

(b) Pentacene is a well-studied organic semiconductor. Look up its meting point.

(c) What values of mobility have been measured for electrons in pentacene? The mobility is a measure of electron flow through the material, which requires both intramolecular flow enabled by the delocalization of the π-band electrons as well as intermolecular flow by hopping. A high degree of crystallization of pentacene is very important to obtain high values of mobility. Explain.

(d) Pentacene is not stable in air. What reaction occurs? How does that affect its feasibility for commercial flexible electronics?

6.3 The simplest OLED display consists of a matrix of rows and columns that allows individual pixels to be addressed. An array of 36 of these pixels is shown below. Each pixel is formed by the intersection of a given cathode electrode strip and a given anode electrode strip.
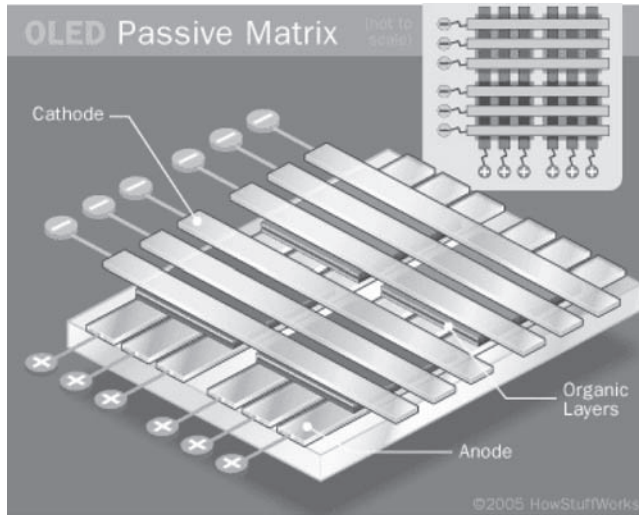
(a) Explain how individual pixels can be controlled by applying appropriate voltages to the row and column electrodes.

(b) An equivalent array of pixels could be made by connecting 36 inorganic LED chips in a matrix of rows and columns. Make a list of manufacturing steps that you might require to fabricate the OLED pixel array and compare this to the list of manufacturing steps needed to fabricate the inorganic LED array.

(c) There are commercial applications of both organic and inorganic pixel arrays. Inorganic LED arrays are used for large displays that are installed outdoors and are viewed from distances of tens of metres above highways or on buildings. OLED arrays are commercially available as cell phone displays. Based on your answer to (b) as well as the requirements of the two applications mentioned, see if you can rationalize the very different applications of the two types of devices.

(d) The largest single market for displays is television. If inorganic LED technology and organic LED technology were being considered for television, what challenges would each technology face to be accepted in this very competitive market? Make a list of strong points and weak points for each technology in the context of the basic requirements of consumer television, which include:

low cost
high brightness
long life
full colour
low power consumption

6.4 A full-colour OLED display contains red, green and blue pixels. In order to generate white light the relative luminance values of these three pixel colours are as follows:

red: 30%
green: 70%
blue: 10%

The brightness specification of the OLED display is measured by the maximum *average* luminance that it can produce in a white screen. Assume that 25% of the display area is red pixels, 25% is green pixels and 25% is blue pixels. A final 25% of the display area is the spaces between pixels, which do not emit light.

(a) For an average white luminance of $500 \, \text{cd m}^{-2}$ find the required luminance values for the red pixels, the green pixels and the blue pixels.

(b) A required specification for this OLED display in terms of lifetime depends on the application. For television, $60\,000$ hours to half luminance is a common requirement. Estimate the lifetime requirement of a cell phone display, and comment on the applicability of current OLED technology for these two applications.

(c) The peak emission wavelengths of the red, green and blue pixels are approximately $615 \, \text{nm}$, $540 \, \text{nm}$ and $470 \, \text{nm}$ respectively. Assuming, for ease of calculation, that the emission in these three colours is monochromatic (light of a single wavelength), find the radiative power per unit area from each pixel for the result of (a). *Hint:* This requires a conversion from photometric to radiometric units, which can be obtained from the plot of luminous efficacy in Chapter 3. You may assume a lambertian source.

(d) For the result of (c) find the relative photon emission rate from each of the three colours in photons per unit area per second. You may assume a lambertian source.

6.5 The structures of OLED devices shown in Figures 6.12 and 6.13 will not yield good contrast because of the high reflectivity of the aluminium rear electrode. In a display application, ambient light will reflect from the rear electrode and be re-emitted. This reflected light may exceed the light emitted by the OLED material and result in poor display contrast. In order to resolve this problem a *circular polarizer* may be placed in front of the OLED. The circular polarizer will prevent the reflection of ambient light off a *specular reflector* such as the aluminium rear electrode. The OLED device will therefore appear black unless it is electrically excited.

(a) Find reference material on polarizers and briefly explain both linear polarization and circular polarization of electromagnetic radiation. How does the circular polarizer prevent the re-emission of ambient light from a specular reflector placed behind the polarizer? *Hint:* Ambient light must first pass through the circular polarizer in one direction before reflecting off the specular reflector and leaving through the circular polarizer in the opposite direction.

(b) The circular polarizer will also attenuate the desired light emission from the OLED. If OLED electroluminescence is not polarized, calculate the maximum fraction of the generated OLED light that could pass through the circular polarizer.

(c) The OLED devices of Figures 6.12 and 6.13 may be simplified from an optical viewpoint as in the following figure, which shows the light emission from the OLED device with an aluminium rear electrode:
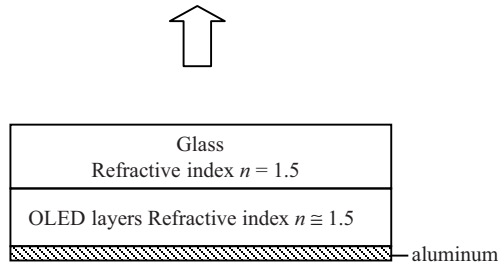
(d) A viewing cone will be formed by considering the critical angle of OLED light emission that is generated in the OLED layer, travels through the glass and is incident on the front surface of the glass. At incident angles higher than the critical angle this light will be reflected back into the device off the front surface of the glass. Find the critical angle relative to the normal of the glass. What fraction of generated light can leave the OLED device? *Hint:* See Section 5.5.

(e) How might the fraction of generated OLED light leaving the device be increased? *Hint:* Use concepts of optical outcoupling in Chapter 5 to discuss this.

6.6 Although the organic solar cell has developed rapidly since 2001, it is well behind inorganic technology in terms of maturity. A major challenge is the stability of organic solar cells. Traditional solar farm installations and rooftop installations demand 20- to 30-year lifetimes with not more than a 20% reduction in performance. This specification will not be easy to meet using organic solar cells. Fortunately there are applications of organic solar cells that require much shorter lifetimes.

(a) Given the inherent flexibility and light weight of organic solar cells, describe some existing applications that take advantage of the special attributes of organic solar cells and that are also less demanding in terms of efficiency and longevity.

(b) Now add some potential *new* applications of organic solar cells to your list. Be creative. New applications are often those that force new technologies to mature quickly and become mainstream since they fill such market niches better than anything else.

6.7 The promise of flexible displays has been a major driver for OLED development. Since OLED active layers as well as OLED cathode materials are highly moisture and oxygen sensitive it is necessary to encapsulate the devices. Flexible polymer sheets are not impermeable since water or oxygen molecules can diffuse between the polymer molecules due to their weak van der Waals intermolecular bonding. Polymer sheets therefore do not offer protection from the atmosphere to OLED layers deposited on them, resulting in rapid degradation. To remedy this, a series of inorganic thin films may be deposited on a polymer sheet with thin organic layers between them to keep them separated. The concept is that the inorganic layers, being impermeable to moisture and oxygen, will render the sheet impermeable. Since there are inevitable cracks in any given

inorganic layer, the use of multiple inorganic layers will force unwanted molecules to take a very tortuous pathway to pass through the sheet as illustrated below:
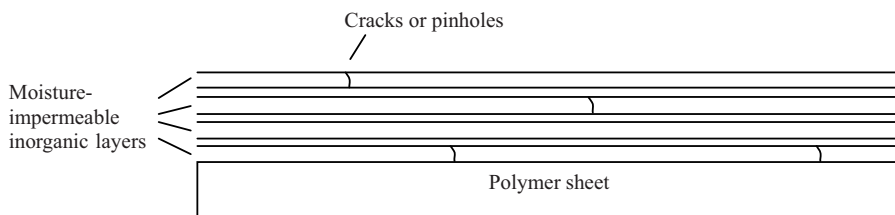
(a) Look up information on multilayer moisture barrier technology for flexible OLED devices. *Hint:* Search using the following keywords: multilayer OLED moisture protection.
(b) Find one or two quoted maximum moisture penetration rates allowable for OLED devices in units of grams per square metre per day.
(c) Flexible OLED devices have not been commercially successful yet. Can you suggest reasons for this?

6.8 Research is underway on organic-inorganic solar cell structures in which the bulk heterojunction has an inorganic acceptor material combined with an organic donor material. For example, the HTL may be solution-processed MEH-PPV and the ETL may be ZnO. The motivation for this is the availability and stability of inorganic electron conductors. An example of such as structure is shown below:
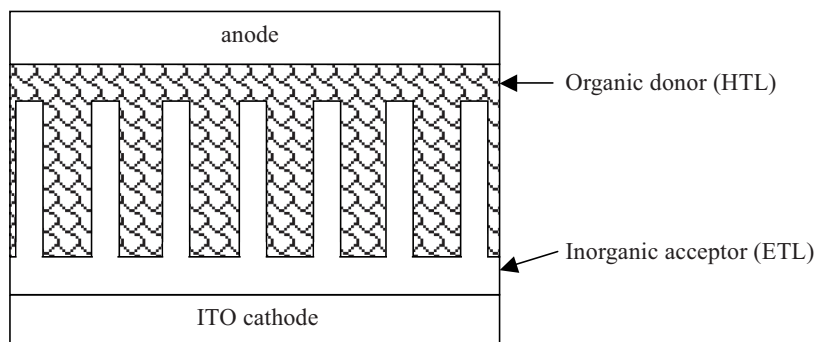
(a) What challenges are there with this approach? *Hint:* Think carefully about what needs to happen right at the inorganic-organic junction, and what could go wrong at the junction.
(b) Search for more information on recent developments in this field. Use the following keywords: organic solar cells; organic/inorganic heterojunctions.
(c) Organic electron-conducting acceptor materials tend to be more moisture and oxygen sensitive than hole-conducting donor materials. Why?
(d) An OLED can also be made using a combination of organic and inorganic materials for the electron-hole injection and recombination processes. Provide and explain an

example. What challenges are inherent in this approach? *Hint:* Use the following keywords: quantum dot light emitting devices.

6.9 A typical OLED display emits light as a lambertian source. This means that the luminance of the OLED display is independent of the angle between the observer and the normal axis to the plane of the OLED display. See Section 5.5.

If the lambertian OLED display provides a luminance of 100 cd m^{-2}, has a light emitting surface area of 100 cm^2 and requires an electrical input power of 1 watt,

(a) Show that the total luminous flux emitted by the display is 3.14 lumens. Hint: Use spherical polar coordinates. Integrate the amount of light emitted over the solid angle range as follows: 1) Obtain the luminous intensity in candelas emitted as a function of angle by multiplying the luminance by the OLED surface area in square meters subtended by the viewer at each viewing angle. 2) To obtain the total luminous flux in lumens (lm), integrate the luminous intensity over the entire range of solid angles that is viewable. Luminous intensity may be expressed as lm Sr^{-1}. Note the definition of luminous flux in Section 3.9. Use spherical polar coordinates.

(b) Find the luminous efficiency of the OLED display in lm W^{-1}. Compare this to luminous efficiency values for inorganic LEDs in Figure 5.1.

Appendix 1: Physical Constants

Boltzmann's constant	$k = 1.38 \times 10^{-23}$ J K^{-1} $= 8.62 \times 10^{-5}$ eV K^{-1}
Electron charge magnitude	$q = 1.6 \times 10^{-19}$ C
Electron mass	$m_e = 9.11 \times 10^{-31}$ kg
Permittivity of free space	$\epsilon_0 = 8.85 \times 10^{-12}$ F m^{-1} $= 8.85 \times 10^{-14}$ F cm^{-1}
Planck's constant	$h = 6.63 \times 10^{-34}$ J s
kT at room temperature	$kT \cong 0.026$ eV
Speed of light	$c = 3.00 \times 10^8$ m s^{-1}

Appendix 2: Properties of Semiconductor Materials

Semiconductor	Energy gap type	Energy gap E_g (eV)	Electron mobility μ_n (cm V^{-1} s^{-1})	Hole mobility μ_p (cm V^{-1} s^{-1})	Relative electron effective mass $\frac{m_n^*}{m}$	Relative hole effective mass $\frac{m_p^*}{m}$	Lattice constant (Å)	Relative dielectric constant ϵ_r	Density (g cm^{-3})
Si	Indirect	1.11	1350	480	1.08	0.56	5.43	11.8	2.33
Ge	Indirect	0.67	3900	1900	0.55	0.37	5.65	16	5.32
SiC	Indirect	2.86	500	–	0.6	1.0	3.08	10.2	3.21
AlP	Indirect	2.45	80	–	0.21	0.15	5.46	9.8	2.40
AlAs	Indirect	2.16	1200	420	2.0	*	5.66	10.9	3.60
AlSb	Indirect	1.6	200	300	0.12	0.98	6.14	11	4.26
GaP	Indirect	2.26	300	150		*	5.45	11.1	4.13
GaAs	Direct	1.43	8500	400	0.067	*	5.65	13.2	5.31
GaN	Direct	3.4	380	–	0.19	0.6	4.5	12.2	6.1
GaSb	Direct	0.7	5000	1000	0.042	*	6.09	15.7	5.61
InP	Direct	1.35	4000	100	0.077	*	5.87	12.4	4.79
InAs	Direct	0.36	22 600	200	0.023	*	6.06	14.6	5.67
InSb	Direct	0.18	100 000	1700	0.014	*	6.48	17.7	5.78
ZnS	Direct	3.6	180	10	0.28	–	5.41	8.9	4.09
ZnSe	Direct	2.7	600	28	0.14	0.6	5.67	9.2	5.65
ZnTe	Direct	2.25	530	100	0.18	0.65	6.10	10.4	5.51
CdS	Direct	2.42	250	15	0.21	0.80	4.14	8.9	4.82
CdSe	Direct	1.73	800	–	0.13	0.45	4.30	10.2	5.81
CdTe	Direct	1.58	1050	100	0.10	0.37	6.48	10.2	6.20

*There is more than one value since both heavy holes and light holes must be considered. The reader is directed to seek more specialized literature on these materials.
Dash indicates value is not known.

Appendix 3: The Boltzmann Distribution Function

The distribution of energies and velocities in an ensemble of interacting atoms or molecules lies at the heart of the quantum distribution functions. In this section we will review the assumptions and the methodology to calculate the probability distribution function of the energies of an ensemble of atoms or molecules that is used as the basis for the calculation of the Fermi–Dirac distribution starting in Section 1.9.

The Boltzmann Distribution Function

Consider a system in thermal equilibrium containing a large number of identical entities. Each entity, which could be an atom or a molecule, will have some energy E with a certain energy distribution function. Energy from each entity can be exchanged with other entities by means of collisions between entities. These collisions are assumed to be elastic, meaning that the total system energy is not changed as a result of energy exchange. The purpose of this section is to find this energy distribution function.

The following summarizes the assumptions that apply to the system of entities:

(a) The system is in thermal equilibrium at temperature T.
(b) Entities in the system can exchange energy by means of perfectly elastic collisions. This means that the total system energy is not altered when a collision takes place.
(c) The total system energy is constant.
(d) All possible divisions of total system energy among the entities in the system occur with equal probability. There is no preferred division of energy among the entities.
(e) The entities are distinguishable. This means that we can identify any given entity and its energy at any time.

Consider an example of this situation. Assume there are four atoms, and each atom can have energy levels of 0, ΔE, $2\Delta E$, $3\Delta E$. Now assume the total system energy is $3\Delta E$. All the possible divisions of energy are as listed in the following cases:

1. 3 atoms at $E = 0$ and 1 atom at $E = 3\Delta E$
2. 2 atoms at $E = 0$, 1 atom at $E = \Delta E$, and 1 atom at $E = 2\Delta E$
3. 1 atom at $E = 0$ and 3 atoms at $E = \Delta E$

These three cases do not cover all the possible divisions of total system energy, however. This is because in case 1 any one of the four atoms could be at energy $E = 3\Delta E$, and case 1 therefore really includes four distinct divisions of energy among the atoms.

In case 2, the atom at $E = 2\Delta E$ could be chosen from any one of the four atoms, leaving three atoms. Any one of the remaining three atoms could now be chosen for $E = \Delta E$. Finally the remaining two atoms would be at $E = 0$. Case 2 therefore results in $4 \times 3 = 12$ distinct divisions of energy among the atoms.

Finally in case 3, any one of the four atoms could be at energy $E = 0$, and case 3 therefore really includes four distinct divisions of energy among the atoms.

Now we can make a list of all 20 divisions and their probabilities of occurrence as shown in Table A3.1. For any one atom, we can now find the energy distribution by looking down the relevant column of the table and counting the number of entries having specific energies. For any one atom:

There is a chance of 1/20 that the atom is in energy state $E = 3\Delta E$.
There is a chance of 3/20 that the atom is in energy state $E = 2\Delta E$.
There is a chance of 6/20 that the atom is in energy state $E = \Delta E$.
There is a chance of 10/20 that the atom is in energy state $E = 0$.

We can plot this energy distribution function as shown in Figure A3.1. The probability of occurrence as a function of energy is the energy distribution function for the system.

Table A3.1   *A list of all possible divisions of total system energy among the atoms in the system. Each division occurs with equal probability*

Arrangement	Atom 1	Atom 2	Atom 3	Atom 4	Probability of occurrence
1	0	0	0	$3\Delta E$	1/20
2	0	0	$3\Delta E$	0	1/20
3	0	$3\Delta E$	0	0	1/20
4	$3\Delta E$	0	0	0	1/20
5	0	0	ΔE	$2\Delta E$	1/20
6	0	ΔE	0	$2\Delta E$	1/20
7	ΔE	0	0	$2\Delta E$	1/20
8	0	0	$2\Delta E$	ΔE	1/20
9	0	ΔE	$2\Delta E$	0	1/20
10	ΔE	0	$2\Delta E$	0	1/20
11	0	$2\Delta E$	0	ΔE	1/20
12	0	$2\Delta E$	ΔE	0	1/20
13	ΔE	$2\Delta E$	0	0	1/20
14	$2\Delta E$	0	0	ΔE	1/20
15	$2\Delta E$	0	ΔE	0	1/20
16	$2\Delta E$	ΔE	0	0	1/20
17	0	ΔE	ΔE	ΔE	1/20
18	ΔE	0	ΔE	ΔE	1/20
19	ΔE	ΔE	0	ΔE	1/20
20	ΔE	ΔE	ΔE	0	1/20

Figure A3.1 *Plot of the probability of occurrence as a function of the energy for any atom in the system*

In Figure A3.2 an exponential function is now superimposed on the same data as shown in Figure A3.1. This exponential function is of the form:

$$p(E) = A \exp\left(-\frac{E}{E_0}\right) \qquad \text{(A1.1)}$$

If we choose suitable parameters A and E_0 we can obtain a reasonable fit to the data.
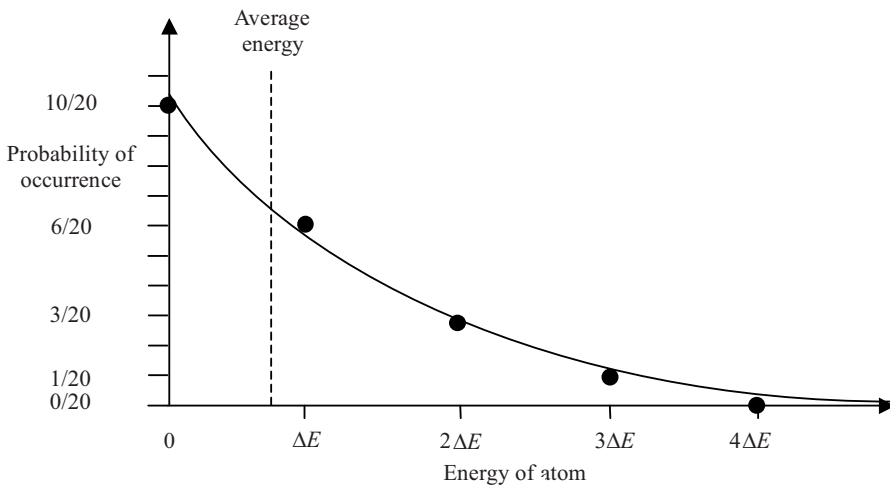
Figure A3.2 *Plot of the probability of occurrence as a function of energy together with an exponential function as an approximation of the data. The average energy at 0.75ΔE is also shown*

If the same procedure using a much larger number of atoms than four is followed, then the number of data points increases rapidly and the ability of an exponential function of the form of Equation A1.1 to represent the data improves and becomes an excellent fit.

The significance of E_0 may now be understood by calculating the *average* energy of any atom in the system. This is done by adding all the possible energies and dividing by the number of divisions of energy. From Table A1.1, the average energy of any one atom is:

$$\frac{(1 \times 3\Delta E) + (3 \times 2\Delta E) + (6 \times \Delta E) + (10 \times 0)}{20} = \frac{15}{20}\Delta E = 0.75\Delta E$$

This makes sense because the total system energy is $3\Delta E$ and there are four atoms each having $0.75\Delta E$. The average energy is shown as a dashed line at $0.75\Delta E$ in Figure A3.2.

It is also possible to calculate the average energy based on the exponential function $p(E)$. This average is calculated again by adding all the possible energies and dividing by the number of divisions of energy; however, since $p(E)$ is a continuous function the addition becomes an integral and we obtain

$$E_{\text{average}} = \frac{\int_0^\infty E p(E) \mathrm{d}E}{\int_0^\infty p(E) \mathrm{d}E} = \frac{\int_0^\infty E \exp\left(-\dfrac{E}{E_0}\right) \mathrm{d}E}{\int_0^\infty \exp\left(-\dfrac{E}{E_0}\right) \mathrm{d}E}$$

The numerator may be integrated by the methods of integration by parts to obtain E_0^2 and the denominator is readily integrated as E_0. Hence we obtain

$$E_{\text{average}} = \frac{E_0^2}{E_0} = E_0$$

Therefore the average energy of one atom in a system with many atoms is E_0.

If the average energy of an atom belonging to a system in thermodynamic equilibrium at temperature T is defined as kT where k is Boltzmann's constant, then we have the famous Boltzmann distribution function

$$p(E) = A \exp\left(-\frac{E}{kT}\right)$$

which is used in Equation 1.20 in the derivation of the Fermi–Dirac distribution function.

Note that we have demonstrated but not *proved* that the exponential function is the correct distribution function in a system with many atoms. This can be proved and the proof is covered in physics textbooks on statistical thermodynamics.

Index