Chapter 12
Semiconductor Lasers

12.1 Introduction

That certain materials behave like a semiconductor is believed to have been first scientifically documented in the work of British scientist Michael Faraday (1791–1867). He discovered in 1833 that the resistance of silver sulfide decreased with increasing temperature in complete contrast to the behavior of metals.[1] These semiconductor materials behave much like insulators at low temperatures but conduct electricity like metals at room or elevated temperatures. It, however, took more than a century before the electrical properties of a semiconductor could be manipulated by American engineer Russell Shoemaker Ohl (1898–1987) to build the p-n junction, which is now ubiquitous in electronic and optoelectronic devices. The p-n junction when suitably biased behaves like a diode and is regarded as the very first semiconductor electronic device. The transistor, a variant of the p-n junction, brought a revolution in electronics and, no wonder, fetched the 1958 Nobel Prize in Physics to its 1948 inventors William Bradford Shockley (1910–1989), John Bardeen (1908–1991), and Walter Houser Brattain (1902–1987). Bardeen incidentally became the only person to date to get the Nobel Prize in Physics twice; in 1972, he and his collaborators Leon N Cooper and John Robert Schrieffer were awarded the coveted prize for developing the BCS theory to explain conventional superconductivity. The credit of inventing the light emitting diode (LED) goes to Nick Holonyak[2] (1928–2022), an American engineer, who, in 1962, while working with General Electric, succeeded in generating spontaneous red light from a forward

[1] The resistivity of metals increases with temperature.

[2] In 2006, the American Institute of Physics announced five most important papers in the journals published by it since it was founded 75 years ago. One of these five papers happened to be the *Applied Physics Letters* paper that Holonyak coauthored with S. F. Bevacqua to report the creation of the first LED in 1962.

D. J. Biswas, *A Beginner's Guide to Lasers and Their Applications, Part 1*, Undergraduate Lecture Notes in Physics,
https://doi.org/10.1007/978-3-031-24330-1_12

biased p-n junction of GaAs semiconductor. This operation of LED provided a major impetus to build a semiconductor diode laser, and by the end of that year, Robert N. Hall (1919–2016), an American physicist, while working at General Electric, Schenectady, New York, generated coherent light from a gallium arsenide diode. The race to build the diode laser was so intense that almost simultaneously the groups working at the MIT Lincoln Laboratory, Massachusetts, and the IBM T.J. Watson Research Center also reported lasing from GaAs diode. Erstwhile Soviet Union was not behind either as in early 1963 Nikolay Basov and his associates developed a gallium arsenide diode laser at the Lebedev Physical Institute, Moscow. That in the following year he was awarded the Nobel Prize in Physics for his pioneering contribution to the invention of laser is now a history. The inbuilt tiny size of the semiconductor laser enabling its ready integration with other devices primarily led to an explosion of research activity across the world to perfect the diode laser technology by exploiting the unique properties that a semiconductor is known to offer. This has led to the emergence of diode lasers as the undisputed leader in the world of lasers; the readers may refer to the second volume of this book to catch a glimpse on the seemingly endless applications of these lasers covering almost all branches of science and technology. As a matter of fact, to say that today diode lasers have made profound impact on our day-to-day life will not be an overstatement; they are present everywhere – CD players, DVD players, barcode readers, laser pointers, cellphones, HD TV – and most importantly they are central to the ever-expanding global communication network making possible the hassle-free long distance trans-pacific and transatlantic video/audio calls.

To be able to appreciate the workings of diode lasers and their diversity and versatility, it is imperative to gain insight into the physics of semiconductors. An intuitive approach will be followed here to elucidate the basics of semiconductors prerequisite for an easy understanding of the working of all variants of diode lasers.

12.2 Basics of Semiconductor Physics

The conductivity of a material is essentially governed by the mobility of its constituent electrons, in particular the ones that are farthest from the nucleus of the respective atoms. It is worthwhile in this context to take another look at Bohr's atomic theory, introduced in Chap. 3, which pictures an atom as its electrons revolving around the positively charged nucleus in specific orbits with rapidly increasing radii. Electrons residing in the outermost orbits are the most loosely bound as they are attracted by the nucleus with the least force. To determine the energy levels of such electrons of the isolated atoms when they are brought closer and closer to form a solid, we consider the simplest case of two hydrogen atoms separated by a distance of R (Fig. 12.1). As long as R far exceeds $2r_0$ where r_0 is the radius of the ground state Bohr orbit, there is no appreciable overlap of the wave

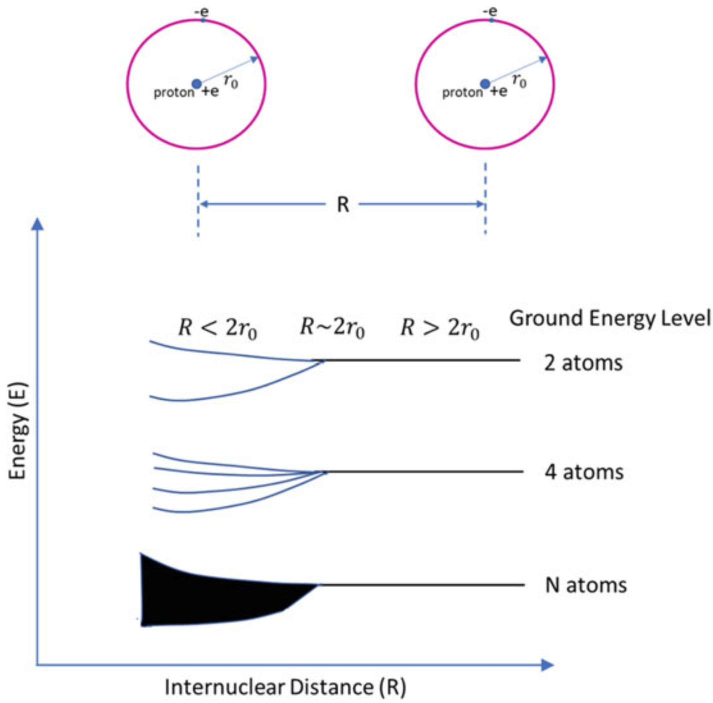


Fig. 12.1 Energy level as a function of interatomic distance. When two hydrogen atoms come close together, the ground state energy level splits into 2 due to the overlap of the corresponding electronic wave functions. Generally speaking, four atoms would split the original level into 4. In reality, there are countless atoms in a solid that would essentially form an energy band

functions[3] of the two atoms, and consequently the discrete ground energy level remains the same as that of the isolated atoms. However, as R approaches $2r_0$, the wave functions begin to overlap, and each electron orbits around both nuclei. Clearly, the wave function of each electron can be approximated as the combination of the wave function of the isolated atoms with both the same and opposite signs. The situation is represented in the illustrations of Fig. 12.2. The two wave functions essentially correspond to two different energy values. This is equivalent to saying that when the two hydrogen atoms are close enough, the ground state energy level splits into two. If there were four atoms instead of two, the energy level would then split into four as the number of new levels equals the number of interacting atoms. Generally speaking, if N atoms are brought together, the energy level would split into N separate levels. In a solid, an exceptionally large number of atoms are packed in a tiny volume, and consequently the levels come so close to each other that they essentially form an energy band, as seen in Fig. 12.1. It needs to be emphasized here that the formation of such energy bands is possible primarily for valence electrons

[3]The description of wave function is a little elaborative and is appended at the end of this chapter.

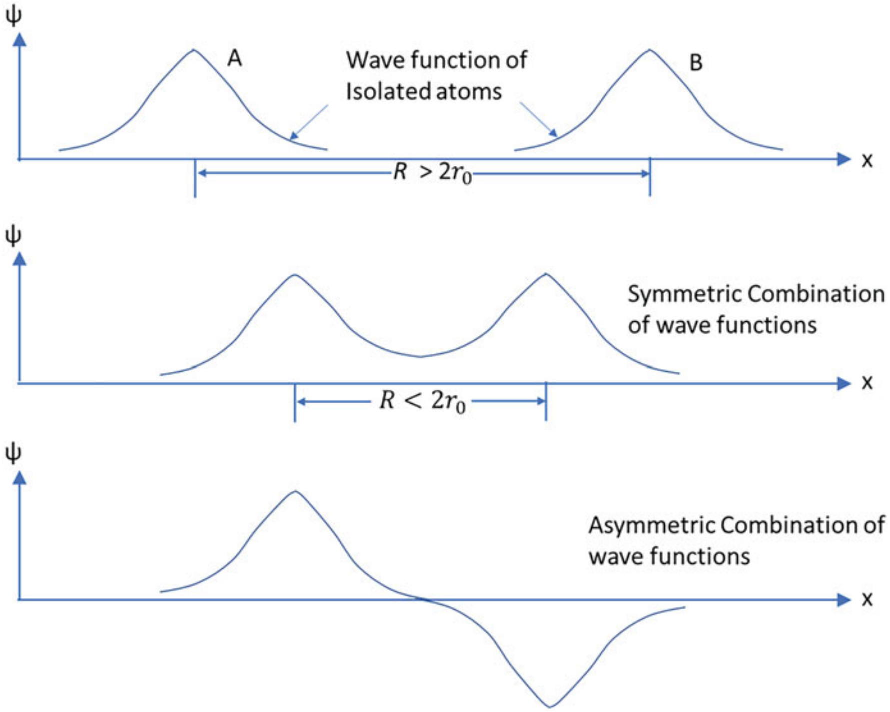


Fig. 12.2 Electronic wave function of two isolated hydrogen atoms (top trace). For small internuclear separation R, the isolated wave functions combine with both same (middle trace) and opposite (lower trace) signs

that are relatively loosely bound to the nucleus as they reside in the outermost shell. Clearly, wave functions of the inner electrons belonging to different atoms forming a solid exhibit very little overlap. We know from Chap. 4 that the shells are denoted by the principal quantum number n. It may be noted here that a shell with a principal quantum number n has an equal number of subshells denoted by orbital quantum number l that can possess integer values 0, 1, 2, 3, ...$(n\text{-}1)$. The maximum number of electrons that can reside in a subshell of orbital quantum number l is $2(2\ l + 1)$. Furthermore, the subshells with $l = 0, 1, 2, 3 \ldots$ are, respectively, denoted as s, p, d, f, etc. Therefore, the maximum occupancy of electrons in s, p, d, f. . .subshells are, respectively, 2, 6, 10, 14, and so on. In order to appreciate the bearing of the formation of the energy band by the valence electrons on the conductivity of a solid, we consider the typical case of a sodium atom that has an atomic number of 11. Each sodium atom thus has 11 electrons, and their distribution among the various shells and subshells is shown in Fig. 12.3. As seen, the shells with principal quantum numbers 1 and 2 are full, and the outermost subshell ($n = 3$, $l = 0$ and accordingly denoted as $3\ s$) is only half full, as only one electron resides here, while the occupancy of the s-subshell is two. This being the valence electron, when sodium atoms get closer to form a solid, the $3\ s$ energy level will be broadened to form an

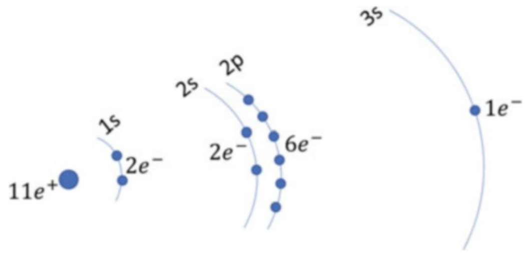


Fig. 12.3 Of the 11 electrons that an atom of sodium possesses, 10 fill all the available slots in the first and second shells, and the 11th electron therefore resides in the l = 0 subshell of the third shell and thus is represented as 3s1

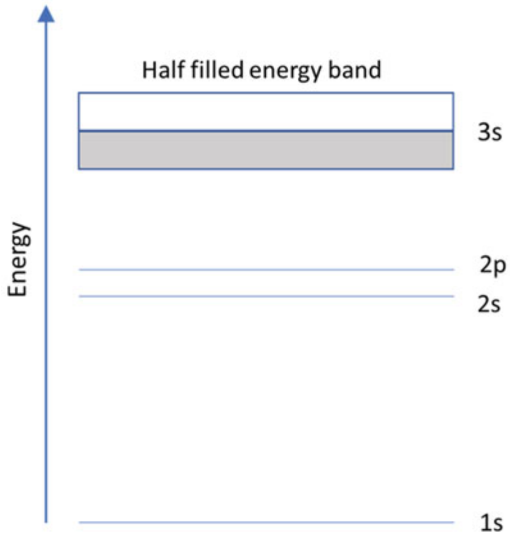


Fig. 12.4 As there is only one valence electron, the '3s' energy band in solid sodium is therefore only half filled

energy band. If *N* atoms have joined together to form the *3 s* band, then it can hold a total of 2 *N* electrons. As only *N* electrons reside in the 3 *s* band, it is therefore only half filled by electrons. The inner levels, viz., *2p*, *2 s*, and *1 s*, being closer to the nucleus, show insignificant broadening of the energy levels. This situation has been described in illustration of Fig. 12.4. Even when a modest potential difference is applied across a block of solid sodium, the *3 s* band electrons can readily acquire energy while still remaining in the same band. The ensuing drift of these valence electrons therefore constitutes a net electric current. It is no wonder that sodium is a good conductor of electricity and is accordingly classified as a metal.

In the case of an insulator, on the other hand, the outermost subshell is completely full, and in turn, the valence energy band here is thus fully occupied (Fig. 12.5). The next outer subshell that is completely empty, and has the potential to function as a conduction band, is separated from this valence band by a forbidden energy gap often measuring several electron volts.[4] The average thermal energy under ambient conditions being only about 0.025 eV, the valence electrons are unable to appear in

[4] 1 eV of energy equals 1.6×10^{-19} joule.

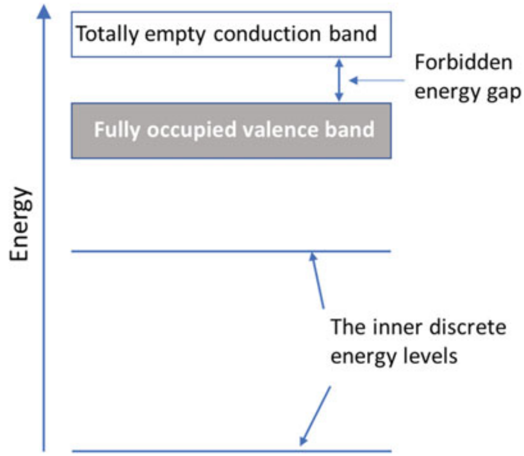


Fig. 12.5 In case of an insulator, a forbidden energy gap separates the fully occupied valence band from the empty conduction band

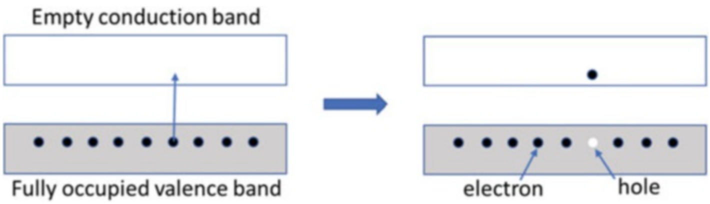


Fig. 12.6 The void created in the valence band when an electron jumps from here to the conduction band is called hole

the conduction band by climbing over the forbidden energy barrier. Nor can they cross such an energy barrier by the application of an electric field of reasonable magnitude that readily sets in a current to flow in a metal. As a matter of fact, an electric field of magnitude to the tune of tens of million V/m is needed for an electron to jump from the valence to conduction band in a typical insulator.

There is yet another kind of substance called a semiconductor that also has a fully occupied valence band like an insulator, but quite unlike it, the empty conduction band here is located only marginally above the valence band; the forbidden energy gap here can be ~1 eV or even less. Such a material behaves like an insulator at low temperature, but at room temperature or above, a small fraction of electrons in the valence band has enough thermal energy to cross over the narrow energy barrier and enter the conduction band. Although small in number, these electrons are still good enough to manifest as a flow of current in the conduction band upon application of a reasonable electric field. The story doesn't end here though, as in contrast to a metal, the semiconductor carries current not only in the conduction band but also in the valence band. The electron, by jumping from the valence to the conduction band, leaves behind an unoccupied state in the valence band called a *hole* (Fig. 12.6). These holes can basically be regarded as a deficit of electrons or equivalently as a positive charge. Just as the motion of electrons in the conduction band results in the

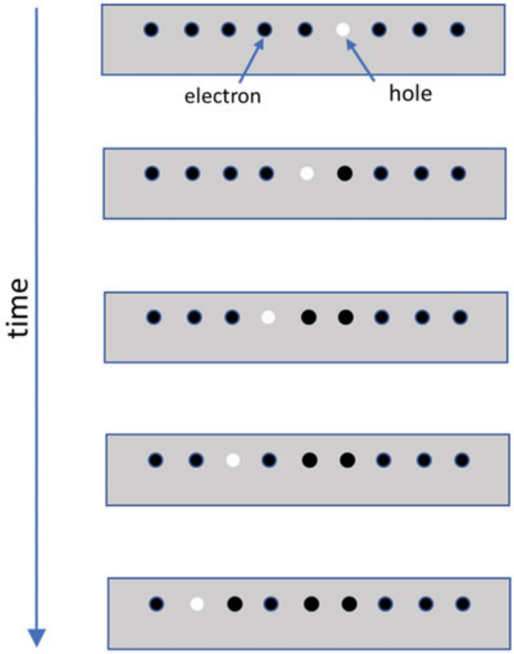


Fig. 12.7 With passage of time as the electrons move from left to right, the hole moves from right to left

flow of a current, the flow of current in the valence band too arises from the motion of holes here. In reality, however, the holes do not move by themselves but appear to be doing so as the valence band electrons basically engage themselves in a game of musical chairs. This point can be readily understood by referring to Fig. 12.7, which essentially represents a collection of snapshots of the motion of the electrons in the valence band captured sequentially in time. Following the creation of a hole in the valence band, one of the neighboring electrons, decided by the polarity of the applied field, moves quickly into it, thereby creating a hole in its original location. As a second electron now moves to fill this new hole, it leaves behind yet again a hole and so on. In a nutshell, therefore, motion of the hole occurs in a direction opposite to that of electrons. Thus, in a semiconductor, the flow of current in the valence band can as well be thought of as originating from the motion of the holes instead of the electrons. This hole-centric approach is also inherently simple to analyze as in the valence band there are fewer holes that are greatly outnumbered by the electrons.

12.3 Impurity Semiconductors

Semiconductor materials exist both in elemental and compound forms. The two most popular element-based semiconductors are silicon and germanium. Some of the molecular variety of semiconductors that find wide applications are GaAs, GaP, InP, etc. (GaAs and InP, as a matter of fact, are the most widely used materials for the operation of semiconductor lasers.) Both these varieties of semiconductors are also called intrinsic semiconductor as the semiconducting properties are inherent to these materials. By virtue of their ability to conduct electricity by both electrons and holes, semiconductors display remarkable properties; viz., the presence of only a trace of impurities can significantly enhance the conductivity of a semiconductor. Semiconductors embedded with such impurities are called *doped* or *extrinsic* semiconductors. While varying the concentration of the dopant in a semiconductor allows ready tailoring of its conductivity, careful selection of the dopant material will allow selecting the type of majority charge carriers. The extrinsic semiconductor for which electrons are the principal charge carriers is called an n-type semiconductor, while the semiconductor with holes as the majority charge carriers is termed a p-type semiconductor. A semiconductor reveals its true potential when a p-type semiconductor and an n-type semiconductor are brought in contact with each other. Interesting effects originating from the passage of the charge carriers through this junction bring about the operation of a variety of semiconductor devices, the p-n junction diode being the simplest of them. In the following section, we shall take a closer look at the underlying physics that allows operation of a semiconductor diode as an emitter of both conventional and laser lights.

12.4 N-Type and P-Type Semiconductors

To understand the construction and operation of these type of semiconductors, let us begin with crystalline silicon as the fundamental intrinsic semiconductor material. A silicon atom has four electrons in its outermost shell, and they, as shown in Fig. 12.8,

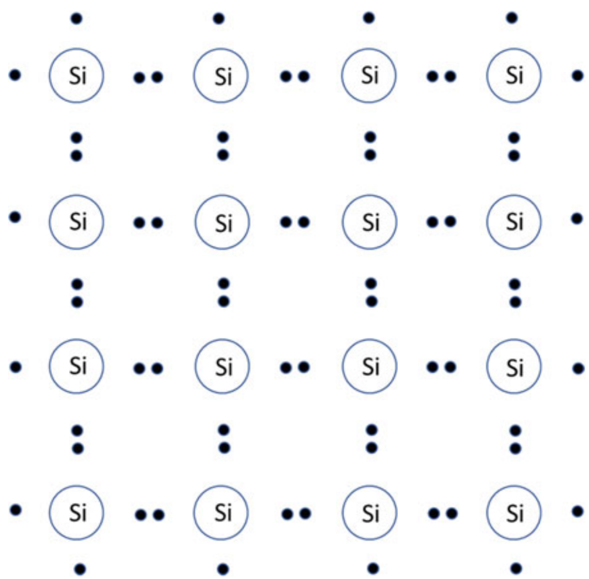


Fig. 12.8 In a crystalline silicon, there is no loose or free electron as all the outer shell electrons are engaged in forming bond with the neighbors

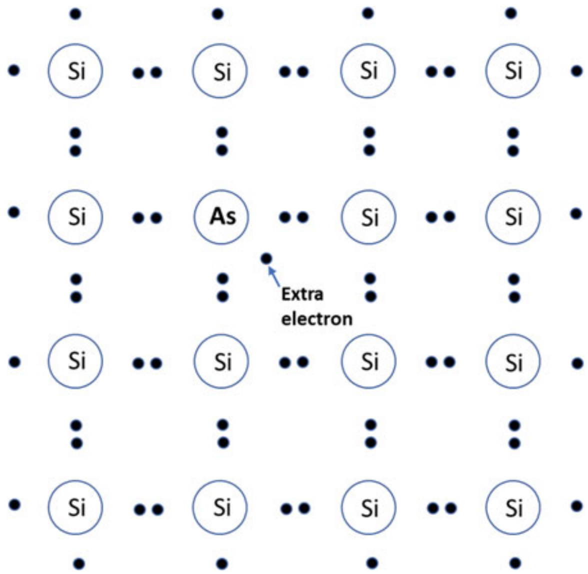


Fig. 12.9 Upon doping, as an arsenic atom displaces a silicon atom in the crystal, the fifth outer shell electron of As with no bonding is now loosely bound. The ability of these electrons to move freely in the crystal dramatically enhances its conductivity

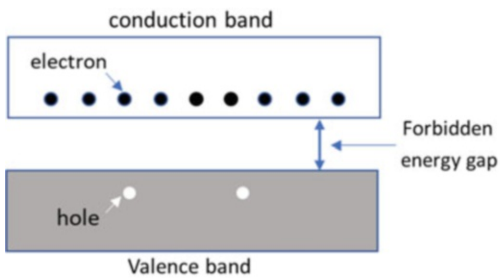


Fig. 12.10 Band structure of an n-type semiconductor revealing clearly that electrons are the majority charge carrier here

are all individually bonded to the electrons of the adjacent atoms while forming the crystal. In the absence of any loose or free electron, the silicon crystal will not conduct electricity unless some electrons from the valence band are excited to the conduction band either thermally or by some other means. However, doping of this silicon crystal with trace quantities of arsenic in a ratio of about 1 in 100 million brings about remarkable changes in its electrical conductivity for the following reason. As a result of doping, the arsenic atoms will replace silicon atoms at a few locations in the crystal. An arsenic atom has five electrons in its outermost shell of which four will be engaged in forming bonds with their nearest *Si* neighbors. The remaining electrons being loosely bound can therefore move freely in the crystal as shown in Fig. 12.9. The arsenic atoms function here as the *donor* of electrons that can easily get into the conduction band. This situation has been schematically illustrated in Fig. 12.10. As seen, the majority of carriers in this extrinsic semiconductor are electrons and hence the name *n*-type semiconductor. The few holes that are present in the valence band owe their origin to the direct excitation of a few

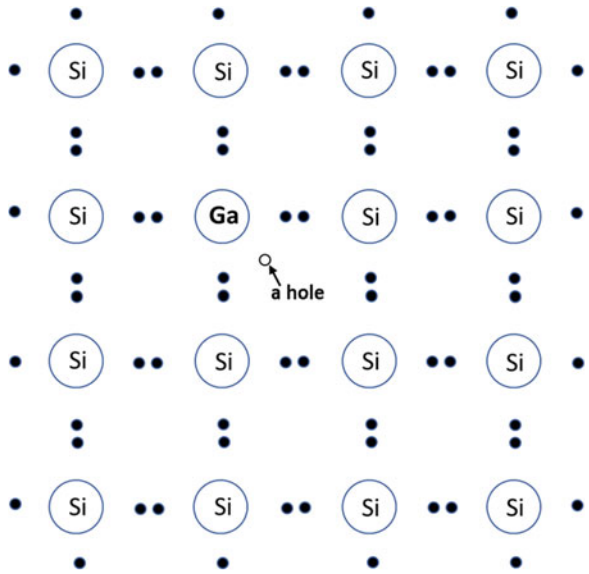


Fig. 12.11 When the silicon crystal is doped with gallium, deficit of electron results in the creation of hole that can move freely in the crystal boosting its conductivity

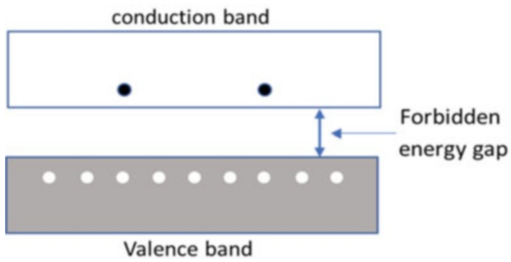


Fig. 12.12 Band structure of a p-type semiconductor revealing clearly that holes are the majority charge carrier here

electrons from the valence to the conduction band. Needless to say, most of the electrons in the conduction band have been donated by the arsenic atoms.

Doping the silicon crystal with trace quantities of gallium atoms, on the other hand, results in a totally different behavior. A gallium atom has only three electrons in the outer shell, and as before, they will replace silicon atoms at a few locations in the crystal. It is not difficult to perceive that as a gallium atom gets electronically bonded with the neighboring silicon atoms, a hole, as depicted in Fig. 12.11, will be formed where the fourth electron should have been present. As already explained, these holes will conduct electricity in the valence band, and the situation has been described in Fig. 12.12. In contrast to the previous case, the majority charge carriers are now holes, and accordingly, this variety of extrinsic semiconductors is termed *p*-type semiconductors. The few electrons present in the conduction band originate from the thermal excitation of some electrons from the valence to the conduction band that, as we know, occurs in an intrinsic semiconductor at room temperature. The gallium atoms thus function here as acceptors of electrons making the movement of the holes possible in the valence band.

12.5 Semiconductor Diodes

Joining together the *p* and *n* junction semiconductors in the manner shown in Fig. 10–38 added a new dimension to the semiconductors paving the way for their device applications. It is worthwhile at this point to take a deeper look at the working of a *p-n* junction. Left to itself, some of the free electrons from the *n*-region will diffuse toward the *p*-region, and likewise, holes from the *p*-side of the junction will diffuse toward the *n*-side (Fig. 12.13a). Electrons exist in the conduction band, while the holes are in the valence band, and as they meet at the junction, the electrons jump into the holes, and both carriers disappear. This leaves in the immediate vicinity of the junction a layer of positive ions in the *n*-side and a layer of negative ions in the *p*-side, as shown in Fig. 12.13b. The charge separation (between the immovable positive and negative ions) sets up a field that opposes any further diffusion of the charge carriers across the junction. Soon, an equilibrium will be established when the opposing field will be strong enough to put a stop on any more diffusion. This results in establishing a thin region in the vicinity of the junction that is devoid of any carriers, called a depleted region, opening up the interesting possibility of using the *p-n* junction as a diode. To understand this, we need to consider two cases. First, we connect the *n*-side to the positive terminal of a battery and the *p*-side to its negative terminal, as shown in Fig. 12.14a. There will be a momentary flow of current as the free electrons from the *n*-region and the free holes from the *p*-region flow away into the battery. This quickly results in the depletion of charge carriers on either side of the junction, thereby causing stoppage of the flow of current through the circuit almost instantaneously. However, the presence of some residual current, which is the reverse current, cannot be ruled out arising from the creation of occasional electron-hole pairs owing to thermal fluctuation that is inevitable. This thus represents operation of the *p-n* junction under reverse bias condition. Next, we reverse the polarity of the battery and the operation here is shown in Fig. 12.14b. The battery now clearly pumps electrons into the *n*-region and removes them from the *p*-region,

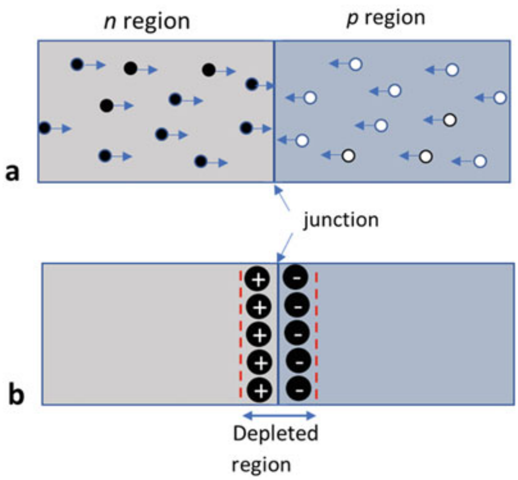


Fig. 12.13 When an n-type semiconductor is joined with the p-type, the loosely bound electrons move from the n to p region. The loosely bound holes, on the other hand, move in the reverse direction (**a**). As the holes and electrons meet at the junction and annihilate each other, a layer of positive and negative ions is formed, respectively, in the n- and p-sides of the junction in the equilibrium. This results in the formation of a thin layer of charge depletion region around the junction (**b**)

Fig. 12.14 The reverse (**a**) and forward (**b**) bias operation of a p-n junction

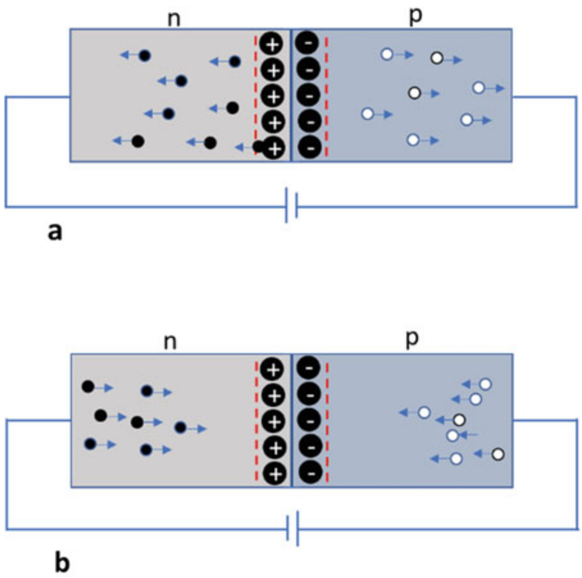
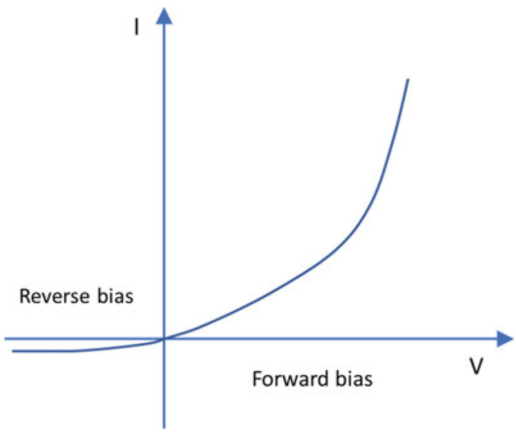


Fig. 12.15 The current as a function of voltage in the operation of a p-n junction diode. The small residual current in case of reverse biasing of the diode is clearly evident



which can also be interpreted as pumping of holes into this region. The holes and electrons continue to meet and annihilate each other in the junction region, maintaining a steady flow of current through the circuit in this configuration known as forward biasing. The higher the battery voltage is, the greater the magnitude of this forward biased current. The voltage-current characteristics of a p-n junction diode operated in both reverse and forward biased conditions are shown in Fig. 12.15. Having acquired this background knowledge of semiconductor physics, we are now in a position to appreciate the working of a semiconductor diode as an emitter of light, both ordinary and coherent.

12.6 Light Emitting Diodes (LED)

Let us in the beginning take a closer look at the creation and annihilation of an electron-hole pair in a semiconductor. We know that a minimum amount of energy equaling the bandgap energy is required to be imparted to a valence band electron for its excitation to the conduction band leaving behind a hole in its original location. Although the omnipresent thermal energy normally facilitates this excitation, it can also be achieved optically provided the photon energy exceeds the bandgap energy. When we say that an electron and hole combine to annihilate each other, what basically happens is that an electron from the conduction band falls into a hole at the valence band and the two nullify each other. As shown in Fig. 12.16, the excess energy that is released when the electron jumps down into the valence band can be emitted as a photon.

Considering the operation of the *p-n* junction, we know that upon forward biasing, it allows a flow of current, while the holes and electrons continue to recombine at the junction and nullify each other. Now recombination is nothing but the conduction band electron literally jumping into the hole in the valence band. If the excess energy of the electron is released as a photon, the diode can then be said to be a light emitter. We can thus say that a *p-n* junction, when forward biased, is capable of emitting light. There is however a catch here. The electron-hole transition is radiative normally for only direct bandgap semiconductors. In a direct bandgap semiconductor, the total momentum of the electron and hole do not change as they

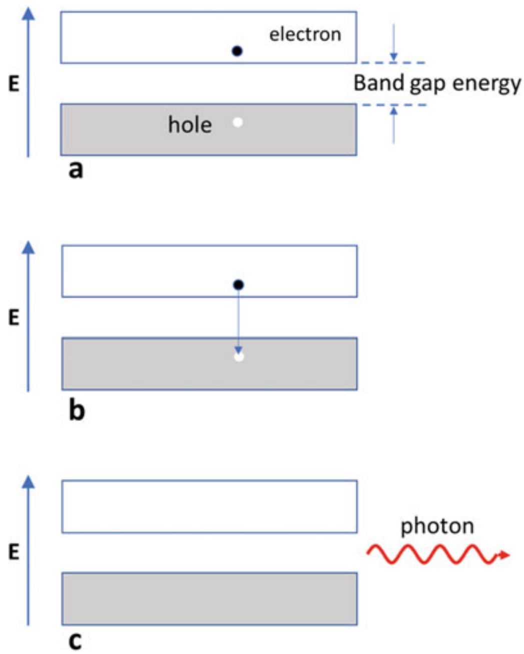


Fig. 12.16 Creation of a photon as an electron and a hole meet and annihilate each other. The electron and the hole originally reside, respectively, in the conduction and valence bands (**a**). As the electron jumps down into the hole (**b**), both disappear and the excess energy carried by the electron is emitted as a photon (**c**)

recombine. The transition only involves a change in energy that is carried by the emitted photon. In an indirect bandgap semiconductor, on the other hand, the total momentum of the electron and hole changes as they recombine. Conservation of momentum is not possible if the excess energy is released only as a photon. The electron needs to pass through an intermediate state to transfer momentum to the crystal lattice, and in turn, the energy is dissipated as heat instead of light. Silicon, the best-known semiconductor material that serves as the building block of miniature electronics, is unfortunately an indirect bandgap semiconductor and thus does not make a good light emitter and hence an LED. Other semiconductors, such as germanium, GaP, AlP, AlAs, and AlSb, also have indirect bandgaps and thus perform poorly as emitters of light. Some of the notable examples of a direct bandgap semiconductor are GaAs, GaSb, InAs, InP, and InSb, of which GaAs has the distinction of emerging as the very first light emitting diode.

12.7 Diode Lasers

Thus, there exists a semiconductor device that is capable of emitting light. The emission here is spontaneous in nature, and the emitted light is therefore ordinary and spreads out in all possible directions. It is not difficult to figure out that a light emitting diode can be converted into a laser by realizing in it the condition of population inversion and simultaneously ensuring the presence of optical feedback. The population inversion can be ensured by appropriately increasing the pumping current so that the rate at which electrons fall into holes and recombine exceeds the rate at which electrons at the valence band can jump into the conduction band by absorbing photons emitted from the recombination. Feedback, as we know, can be provided by placing the device inside an optical cavity. Nature has been kind here, since as the refractive index of semiconductor materials is usually high (e.g., $\mu = 3.6$ for GaAs), the Fresnel reflectivity[5] $\left[\frac{(\mu-1)^2}{(\mu+1)^2}\right]$ is therefore also quite high (~32% for GaAs). Cleaving the end faces perpendicular to the laser axis will therefore allow much of the light created in the junction to be reflected back and forth. This will help meet the threshold condition for lasing in many cases. The configuration of a typical diode laser is schematically illustrated in Fig. 12.17. As it occurs in any other laser, a fraction of the spontaneously emitted photons in the junction that travel along the axis will initiate the stimulated emission causing amplification of light. For low gain lasers or when a unidirectional laser beam is essential, the reflectivity of the faces can be manipulated by appropriately coating them with dielectric layers.

[5]Readers may take another look at Sect. 2.3.1 of Chap. 2 for refreshing their memory on Fresnel reflection.

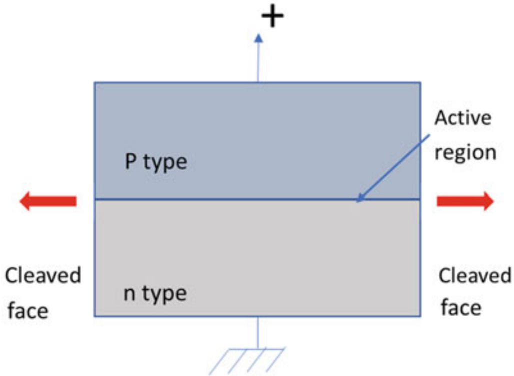


Fig. 12.17 A light emitting diode turns into a laser when the forward biased current exceeds the lasing threshold value. This is a longitudinal or side view of the laser

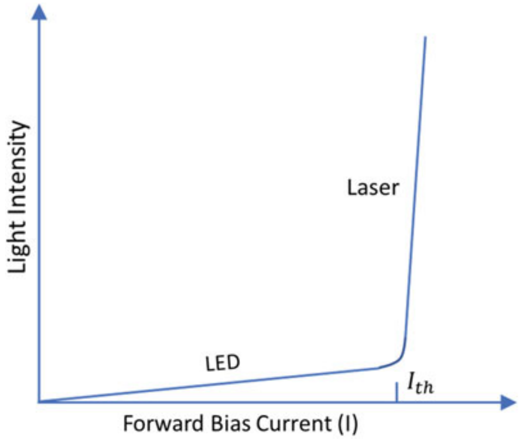
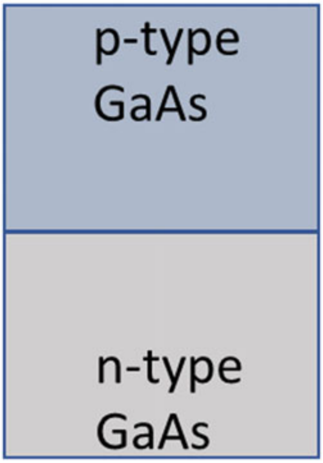


Fig. 12.18 Switching of the operation of a forward biased semiconductor diode from LED into laser with increasing bias current. 'Ith' is the threshold value of current beyond which the diode works as a laser

12.8 Homojunction Diode Lasers

When the diode is made out of the same type of semiconductor material on both sides of the junction, the laser is called a homojunction laser. When the diode is forward biased and the current begins to flow, recombination of electron-hole pairs sets in across the junction resulting in the emergence of spontaneously emitted light. As schematically shown in Fig. 12.18, the intensity of this light gradually increases with the current. This monotonous rise in the intensity of light continues until the current reaches a critical value when it exhibits a dramatic rise, revealing the onset of lasing. The value of the current at which the amplification of light due to stimulated emission begins is called the threshold current for lasing. The requirement of this threshold current is exorbitantly high for a homojunction semiconductor laser for the following reason. Let us consider a GaAs diode, and an appropriately labeled cross-sectional view of the same is shown in Fig. 12.19. Two sides of a GaAs block are doped, the top face as p and the bottom face as n. When the diode is forward biased,

Fig. 12.19 Cross-sectional or end view of a homojunction diode



electrons from the *n*-region and holes from the *p*-region are injected into the junction where they recombine and emit light. As it is GaAs everywhere, this light can easily escape through both the top and bottom faces of the junction. Furthermore, there is nothing to confine the holes and electrons into the junction to ensure a better rate of recombination. These two factors together contribute to the extremely low electro-optic efficiency of a homojunction diode laser. This results in dissipation of excessive heat that can easily damage the diode. To prevent this, it is mandatory to cool the device by placing it in a liquid nitrogen bath (~77° K). However, even with such cooling, they are capable of operating only in the pulsed mode as *cw* operation is impossible to achieve. Although historically significant and easy to understand, these homojunction lasers, with their extra baggage of liquid N_2 bath, are unsuitable for any practical applications.

12.9 Heterojunction Diode Lasers

The limitations of the homojunction lasers can be overcome if both the light and current carriers are confined within the narrow zone of the junction layer, a seemingly uphill task. It is no wonder that the operation of a *cw* diode laser was realized at room temperature more than half a decade after the discovery of lasing in a homojunction diode in the latter half of 1962. Nature was again kind here as a material with a higher bandgap intrinsically possesses a lower refractive index.[6] This

[6]In 1952, Trevor Simpson Moss (1921–), a British physicist, published a book (T.S. Moss, Photoconductivity in the Elements, Academic Press Inc., New York, 1952) during his affiliation to Cambridge University, connecting the electronic and optical properties of a semiconductor by a rule that is known as Moss rule and is given by *Band Gap energy* × *refractive index* = *constant*.

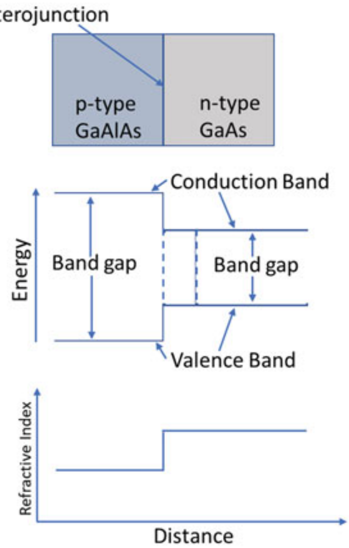


Fig. 12.20 A semiconductor material of higher bandgap has a smaller r.i. and vice versa. A heterojunction therefore neither allows passage of current nor light (when angle of incidence exceeds critical angle) from the lower to higher bandgap side

advantage was exploited in 1968 by Herbert Kroemer (b–1928), a German-American physicist, and Zhores Alferov (1930–2019), a Soviet and Russian physicist, when they succeeded independently in the confinement of both light and current inside the junction layer. The conception of a novel heterojunction diode did the trick here, a masterstroke that killed two flies with one slap. The room temperature operation of the *cw* diode laser eventually fetched Kroemer and Alferov the 2000 Nobel Prize in Physics.

In order to simplify the understanding of the physics of confinement, we consider at this point a single heterojunction diode as shown in the top trace of Fig. 12.20; *p*-type GaAlAs and *n*-type GaAs make a heterojunction with it. GaAlAs is a higher bandgap semiconductor, and the electrons that are injected into the junction from the *n*-side lack energy to cross over into it as GaAs is a lower bandgap semiconductor. This situation is depicted in the middle trace of this figure. This, in turn, results in an improved confinement of the charge carriers in a single heterojunction laser. It readily follows from Moss's rule that GaAs, being a lower bandgap material, will have a refractive index exceeding that of GaAlAs. Consequently, the passage of light from GaAs to GaAlAs will be blocked as the phenomenon of total internal reflection will be operative here. A single heterojunction laser thus allows only partial confinement of current and light and is capable of operating in the pulsed mode at room temperature. The confinement would be much better by going for a double structure heterojunction diode, as shown in Fig. 12.21. As seen, a thin slice of intrinsic GaAs semiconductor is basically sandwiched here by *p*-type GaAlAs on the left and *n*-type GaAlAs on the right. The double-heterojunction diode laser with improved confinement of both current and light within the lasing volume therefore ensures that the majority of the charge carriers recombine here and emit light. As shown in Fig. 12.22, the threshold current for the onset of lasing is therefore remarkably less

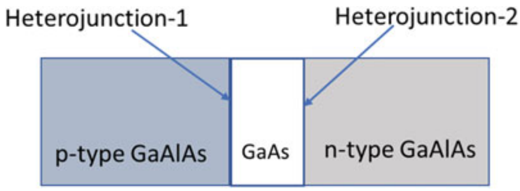


Fig. 12.21 A double-heterojunction diode when appropriately biased is endowed with the capability of cw lasing at room temperature

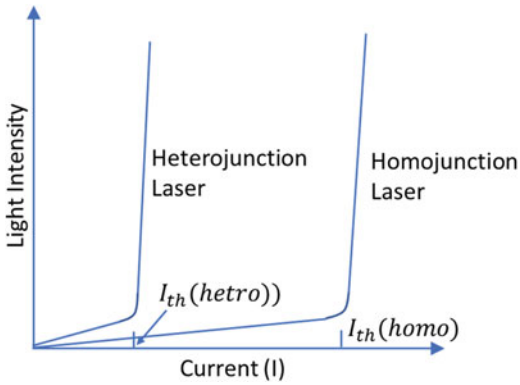


Fig. 12.22 The better confinement of current and light in a heterojunction diode laser lowers the threshold current for lasing significantly. The behavior of a homojunction laser is also shown here for the sake of comparison

than that for the homojunction lasers. This results in a significant improvement in the performance of the laser by establishing increased electro-optic efficiency and reduced heat dissipation. It is not surprising that room temperature *cw* operation is readily achievable from a double-heterojunction laser. As a matter of fact, the semiconductor lasers of today are primarily based on such double-heterostructure design of the diodes.

12.10 Quantum Well Lasers

Double-heterojunction lasers, we now know, offer lower lasing threshold current and higher electro-optic efficiency primarily because they confine the charge carriers in the active layer between the junctions. This enhances the probability of electrons and holes recombining and emitting light that undergoes total internal reflection within this double heterostructure and cannot readily escape. Quite naturally, therefore, the thinner the active layer is, the more likely the recombination, and the better the laser performs. With the advancement of layer deposition technology, it is now possible to fabricate semiconductor layers as thin as tens of nanometers. At such a small thickness that is comparable to the de Broglie wavelength of the electron, the quantum size effect sets in and profoundly impacts the operation of the diode lasers. The electrons now behave like a particle in a quantum well, resulting in quantization of the energy levels as shown in Fig. 12.23. This is akin to the case of a particle in a box leading to energy quantization and has been elaborated in Chap. 3. Exactly the

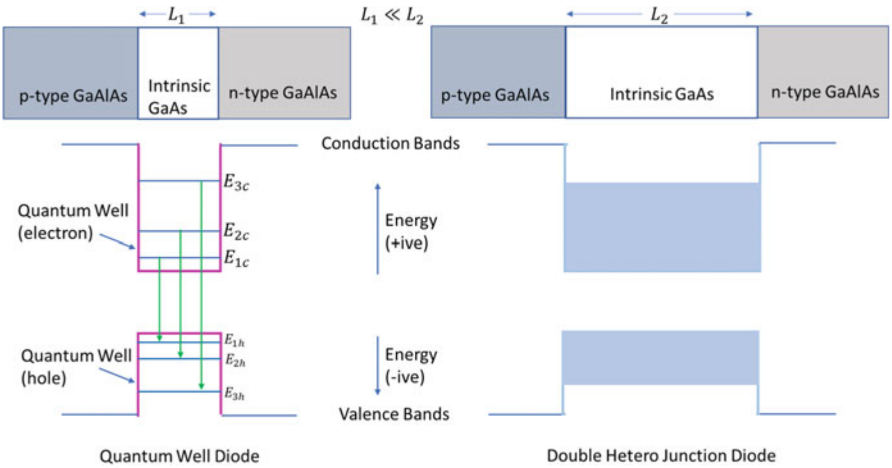


Fig. 12.23 When the thickness of the junction layer approaches the de Broglie wavelength of the electron, the energy levels are quantized for both electrons in the conduction band and holes in the valence band. The conduction and valence energy bands in case of a double heterojunction are also shown here. The thickness of the active layer has been exaggerated to clearly distinguish the two diodes

same happens to the energy levels of the holes in the valence band as well but only in a direction opposite to that in the conduction band forming an upside down well. Holes being heavier than the electrons, energy levels are more closely packed in the case of holes. Only a few low-lying energy levels usually lie inside the quantum wells, and the ones that lie outside bear no consequence in the context of quantum well lasers as the electron or hole with such energies will essentially escape the well. We know from Eq. 3.4 , $E_n \propto \frac{n^2}{mL^2}$, where n, L, and m are the level quantum number, thickness of the active layer, and mass of the charge carrier, respectively. The quantized values of energy can therefore be appropriately tailored to suit the requirement of a particular application by adjusting the layer thickness. The allowed transitions are those for which the n value does not change, and the same are also shown in this figure. With reducing L, the energy levels move away from each other, while with increasing L, the levels approach each other. When L far exceeds the de Broglie wavelength of the electron, the levels become so closely packed as to turn into an energy band, the signature of the double-heterojunction diode, also shown on the right side of this figure for the sake of comparison.

The operation of a quantum well laser is illustrated in Fig. 12.24. A slice of GaAs intrinsic semiconductor sandwiched between a p-type and an n-type GaAlAs semiconductors forms the active layer of the quantum well diode considered here. The extreme narrowness of the active layer allows only a lone energy level to reside inside both the conduction and valence band quantum wells. When the diode is forward biased, electrons and holes injected into the n- and p-regions of the diode by the battery appear in the energy levels of the respective quantum wells. The electrons and holes can now recombine to produce light with remarkable ease in the ultrathin

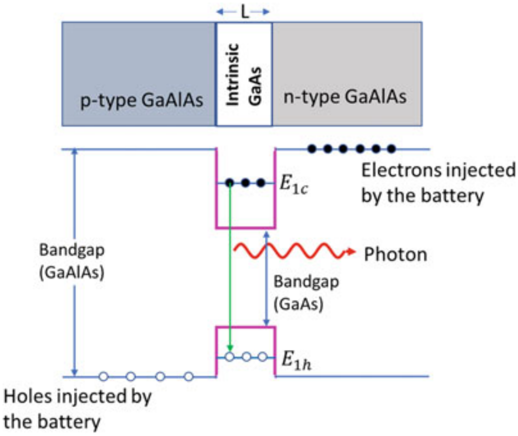


Fig. 12.24 Schematic representation of the working of a quantum well laser

active layer. The inherent high electro-optic efficiency of a quantum well laser allows realization of population inversion and, in turn, lasing at a modest value of the bias current.

Charles H. Henry (1937–2016), an American physicist, conceived the idea of discretizing the energy levels of semiconducting electrons by confining them inside a quantum well in late 1972, while he was heading the Semiconductor Electronics Research Department of Bell Labs, Murray Hill, New Jersey. Later, in 1974, in association with Raymond Dingle, a physicist from the same department, he operated the first quantum well laser[7] and promptly secured a US patent entitled "Quantum Effects in Heterostructure Lasers." The very thin heterostructure, the heart of this device, was fabricated by W. Wiegmann using the technique of molecular-beam epitaxy.[8] Although invented at Bell Laboratories, the name "quantum well laser" was coined by Nick Holonyak, the inventor of LED, and coworkers at the University of Illinois, Urbana-Champaign.

A quantum well laser offers a number of distinct advantages. The first and the foremost is the ease with which the lasing wavelength can be tuned by changing the thickness of the active layer. This can be understood by taking another look at Fig. 12.23. The wavelength of the photon emitted from the radiative transition between energy levels E_{1c} and E_{1h} is essentially governed by the energy difference between these two levels. As the energy quantization and the thickness of the layer are interconnected, the wavelength of the emitted photon will also depend on the layer thickness. More precisely, the wavelength decreases with the reduction in layer thickness and vice versa. To effect such a change in thickness, even on a minute

[7]To know the story of the invention of the quantum well lasers in the inventor's own word, the readers may refer to the foreword that C. H. Henry wrote for the book *Quantum Well Lasers* ed. by Peter S. Zory Jr.

[8]Molecular-beam epitaxy (MBE) is a process for thin film deposition of single crystals. The ultrahigh vacuum and extremely low deposition rate adopted here allow exquisite control over material purity and layer thickness, respectively.

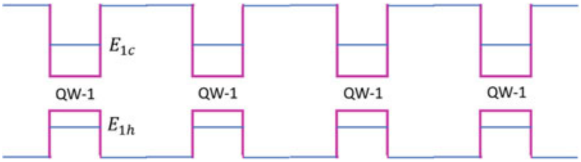


Fig. 12.25 Schematic representation of stacking four quantum wells together to fabricate a multi QW laser

scale, is quite straightforward in practice, e.g., in the MBE, which is widely used in the manufacturing industry of semiconductor devices, the layer thickness can be precisely maneuvered by controlling the thin film deposition time. In contrast, wavelength tuning, albeit discretely, in a typical diode laser is possible only by changing the energy bandgap that mandatorily calls for changing the layer composition, a more complex and involved procedure nevertheless.

As in the greatly reduced volume of the active layer electrons and holes can find each other and recombine with remarkable ease, a quantum well laser thus requires fewer holes and electrons to reach the lasing threshold compared to double-heterojunction lasers. This is the primary reason as to why a QW laser can perform at a high electro-optic efficiency even though the reduced active volume greatly lowers the mode filling factor.[9] The small active volume, however, limits the maximum optical power that can be extracted from a single QW laser. This disadvantage can be overcome by stacking in parallel a few identical quantum well heterojunction structures with a separation judiciously chosen so that they are optically coupled but electronically noninteracting. The nonoverlapping electron wave functions ensure that the energy quantization of individual heterostructures remains unaffected giving rise to multiple quantum well structures, as shown in Fig. 12.25. The strong optical coupling between the individual active layers in conjunction to their integration within a single optical cavity ensures, on the other hand, an output beam of good spatial quality. In summary, the laser now generates an output beam of higher power in accordance with the increased active volume as a result of stacking multiple quantum well structures but with seemingly no ill effect on its spatial quality.

12.11 Quantum Cascade Lasers

Unlike typical semiconductor lasers wherein light originates from inter-band transitions involving the recombination of electrons and holes, a quantum cascade laser operates on intra-band sublevels involving only one type of carrier, viz., an electron. To drive this point home, the emission of a QC laser has been illustrated in Fig. 12.26 alongside that of a QW laser. Although the possibility of a quantum cascade laser

[9] In the operation of a laser, the oscillating mode usually does not have a complete overlap with the active volume of the lasing medium. The fraction of the gain volume that the mode utilizes for its growth is called the “mode filling factor.”

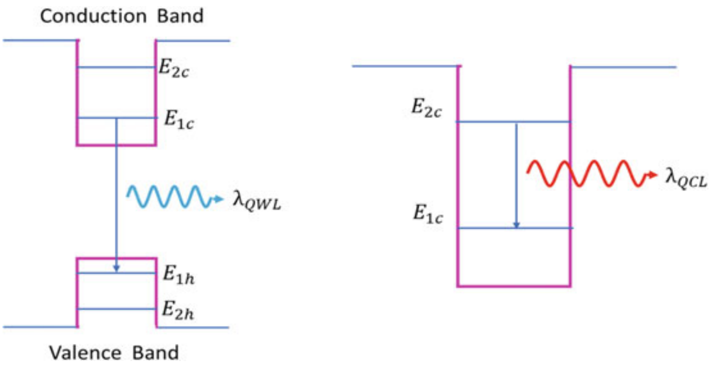


Fig. 12.26 Quantum cascade laser is purely of intra-band origin where the electrons make a transition between the discrete energy levels within the conduction band quantum well (right trace). In case of a quantum well laser, on the other hand, the emission originates from the electron-hole recombination across the material bandgap (left trace)

was conceptualized in 1971, it took more than two decades before such a laser was operated for the first time at the Bell Laboratories, Murray Hill, New Jersey, by Alfred Y. Cho (1937–) and Federico Capasso (1949) and coworkers. As occurs in the case of QW lasers, here too an ultrathin active layer leads to quantum confinement of the electrons resulting in the quantization of their energy levels. The spacing between these levels, as we know, depends on the width of the active layer and can thus be altered by varying the same. The electrons can jump from a higher energy level to a lower energy level releasing the excess energy as a photon. The QC lasers are capable of emitting light of much longer wavelength than achievable for a QW laser, where the lasing transition occurs between widely separated energy levels as one lies in the conduction band QW and the other lies in the valence band QW. These lasers are therefore capable of emitting on the mid- and far-infrared regions of the e-m spectra with wavelengths ranging from several micrometers to hundreds of micrometers and well into the THz region.

As the name suggests, the operation of a QC laser is realized by laying a series of quantum wells that are appropriately biased to essentially form an energy staircase. As the electron tunnels[10] resonantly from one quantum well to the next and travels down the energy staircase, much like a marble rolling down the stair, it emits a photon every time it cascades down an energy step. The situation is schematically described in the illustration of Fig. 12.27. As seen, a single electron, upon its injection into the gain region, is made to emit multiple photons. The number of wells to be coupled is essentially governed by the requirement of the optical power and can vary from a few to tens to even a hundred depending on the applications.

[10]The quantum tunneling effect is the ability of a particle to penetrate a potential barrier with an energy less than the height of the barrier. Exponential decay of the magnitude of the wave function of the particle allows its quantum mechanical penetration through the barrier.

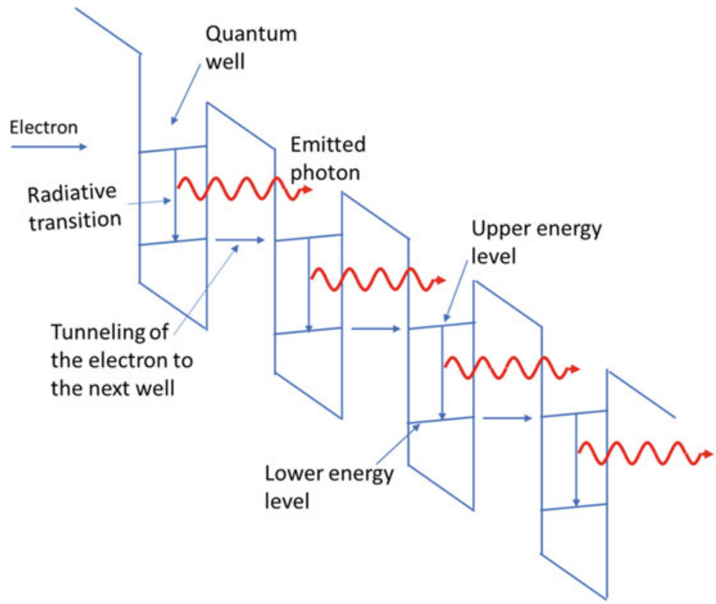


Fig. 12.27 A schematic representation of the principle of working of a QC laser. A judicious combination of the layer thickness and the applied electric field makes possible for the electron to resonantly tunnel to the next well. Consequently, a single electron is able to manufacture multiple photons as it cascades down the energy steps

The inbuilt capability of readily tunable long wavelength operation has made these lasers particularly suitable for trace gas analysis and chemical sensing applications, and the same will be elaborated in volume II of this book.

12.12 Edge and Surface Emitting Diode Lasers

So long our focus has been largely on the variety of compositions and configurations of the heterojunction diode and its active layer with no special emphasis or consideration on the arrangement of the optical cavity. We now turn our attention to the organizations of the optical cavity to determine their bearing on the overall performance of the heterojunction lasers. The early work relied on the cavity with longitudinal geometry that offered two inherent advantages: (1) The active length here equals the length of the junction layer, and (2) the two cleaved end facets with intrinsically high Fresnel reflectivity can perform the role of the cavity mirrors. A laser based on such a cavity would thus emit through its ends and is therefore termed as an edge emitting laser. A typical edge emitting laser with GaAs as the base semiconductor material has been schematically illustrated in Fig. 12.28. As seen, the length and width of the *n*-GaAs substrate are approximately 300 and 200 μm,

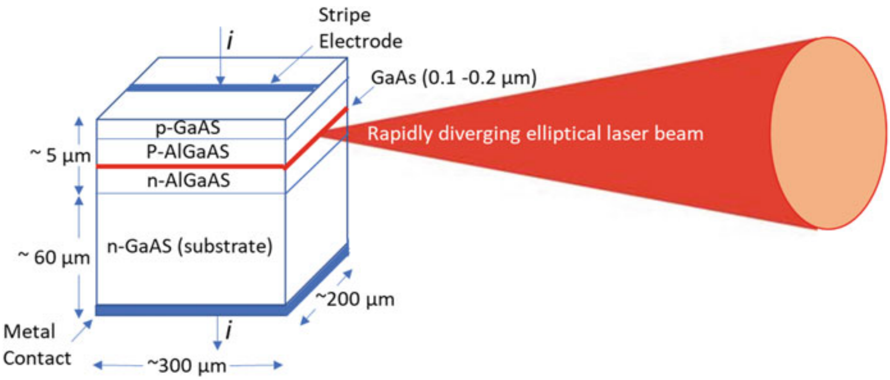


Fig. 12.28 The schematic illustration of an edge emitting double heterojunction diode laser. The ultrathin active layer results in the generation of a highly diverging beam of light that is elliptical in shape

respectively, while the GaAs junction layer is barely 0.1 to 0.2 μm thick. The *p*-AlGaAs and *n*-AlGaAs layers lying on either side of GaAs constitute the *p*-*n* junction. In order to establish a proper ohmic contact, the ~1-μm-thick top layer is made out of *p*-type GaAs. To restrict the current flow through a narrow region to increase the density of charge carriers, the electrical connection is taken by laying a stripe electrode atop. AlGaAs being a higher bandgap material than GaAs, there will be, as we know, an automatic confinement of both the charge carriers and light within the active volume. Having gone through the phenomenon of the diffraction of light in Chap. 2, we now know that the divergence of a beam of light is primarily governed by its size at the source – the narrower the source, the higher its divergence and vice versa. The exceptionally thin active layer of the diode laser (that usually lies within 0.1 to 0.2 μm and is considerably lower than the wavelength of emission) therefore makes the beam emerging through it extremely divergent. As the active zone is smaller in the vertical direction than in the horizontal direction, the divergent beam is thus essentially elliptical in shape. Although the beam can be corrected externally to some extent, this strong diverging component intrinsic to the edge emitting diode laser nevertheless poses a major challenge.

This limitation of edge emitting lasers can be readily overcome by configuring the cavity in the transverse direction, as it allows the ready advantage of greatly enlarging the emitting area of the laser. Popular as VCSEL (vertical cavity surface emitting laser), a beam with remarkably improved spatial quality, which is both round and much less divergent, emerges from such a laser. The arrangement of a VCSEL has been conceptually depicted in the illustration of Fig. 12.29. As can be readily appreciated, this advantage comes at the expense of the gain length, which is now barely a fraction of a micrometer equaling the thickness of the active layer. This greatly reduced active length can be offset by accordingly reducing the round-trip cavity loss. The moderately high reflectivity that the semiconductor materials are known to possess and comes in handy for edge emitting lasers would not thus suffice

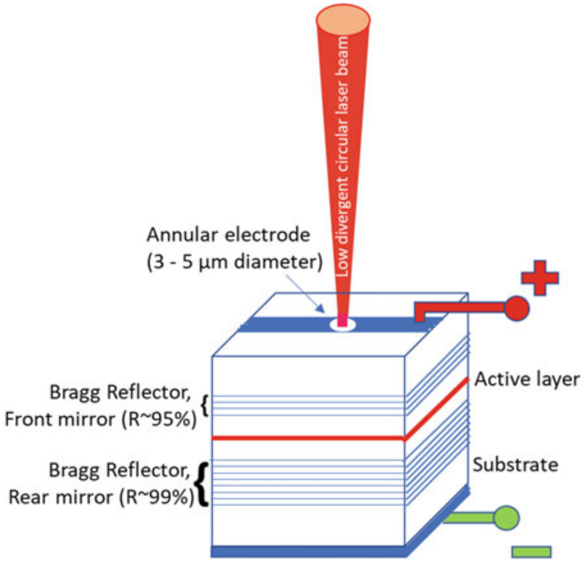


Fig. 12.29 Conceptual depiction of a surface emitting double heterojunction diode laser. Participation of a significantly larger and circular cross section of the active layer into the process of lasing results in the emission of a less divergent and round output beam

here. The use of Bragg reflectors,[11] which can offer remarkably high reflectivity when appropriately designed, has emerged as the most attractive choice to effect lasing in these vertical cavity lasers. As seen, the illustration of Fig. 12.29 makes use of Bragg mirrors of reflectivity ~99% and ~ 95% as rear and front mirrors, respectively, to form the resonator cavity. The top electrode is made annular to allow emergence of the low divergent circular output beam. VCSELs are inherently low-power devices and usually generate less power than edge emitting lasers as the active volume where recombination of charge particles takes place is much too small. These lasers, with their very low lasing current threshold and satisfactory beam quality, are therefore an attractive choice for low-power applications.

12.13 Laser Diode Arrays: From Watts to Kilowatts

The tiny physical size of the active layer in a diode laser essentially limits its power generation capability to a modest level, which may be adequate for a host of the aforementioned applications. However, for applications requiring higher power such as pumping high-power lasers or a variety of material processing and medical applications, power scaling is possible by combining the output of many single emitter devices. The number of diodes to be added depends on the requirement of power, and they can be stacked in a linear array often called a diode bar. To meet the

[11] A Bragg reflector essentially is a mirror structure which consists of a stack of alternating high and low index layers of two different optical materials.

increasing demand of power, the diode bars can be stacked side by side as well as atop each other, forming a regular matrix of emitters with the capability of generating continuous power exceeding a kilowatt. Needless to say, provision of adequate cooling by way of fabricating microchannels within the entire structure is a prerequisite for the working of such a high-power device called a diode array.

12.14 Appendix

Wave Function: In 1801, Thomas Young performed the famed double slit experiment (Chap. 2, Section 2.4) that reestablished the wave nature of light. Light was found to reach a point behind the slits that would be out of bounds if light were to behave here as particles. More than a century and quarter later, the interference of electron was experimentally demonstrated by Clinton Davisson and Lester Germer in 1927 validating the wave nature of microscopic particles, soon after the conceptualization of wave-particle duality in 1924 by Luis de Broglie. It is therefore not surprising that Newtonian mechanics is not adequate to describe the dynamics of a microscopic particle, such as an electron, and this is where quantum mechanics comes in handy. The electron behavior, assuming it to be travelling in a free space, can now be satisfactorily described by Schrodinger's wave equation

$$i\hbar\frac{\partial\Psi(\mathrm{x},\mathrm{t})}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2\psi(x,t)}{\partial x^2}, \text{the quantum equivalent of Newton's law of motion} \quad (12.1)$$

where $\hbar$ is the reduced Planck's constant and m is the mass of electron. More importantly, the function ψ(x,t) is identified as the electron's wave function which is complex valued for every position x and for all times t. The wave function of a particle at a given position and time, therefore, only represents the likelihood of finding the particle at that location and time and has no physical significance. By analogy with, say, an acoustic wave, it can be thought to be an expression for the probability amplitude of the de Broglie wave of the electron. The absolute value squared of the wave function ψ, viz., $|\psi|^2$, being real, however, has a physical significance. It represents the probability density of finding a particle in a given region of space and at a given point in time.