

Concepts of ENGINEERING PHYSICS



NARENDRA MATHAKARI



All of physics is either impossible or trivial. It is impossible until you understand it, and then it becomes trivial.

Ernest Rutherford
(Nobel Laureate, 1908)

About the cover page

The cover page shows the photographs of two huge facilities constructed by Physicists for detection of Higg's Boson and the Gravitational waves. Both the facilities are billion dollars projects. As such, year 2012 and 2015 have been unique as regards to Physics. In 1912, Physicists discovered the theoretically proposed Higgs Boson in world's largest accelerator, the Large Hadron Collider (LHC) sitting in a circular tunnel 27 km in circumference! LHC straddles the Swiss and French borders on the outskirts of Geneva. The LHC is designed to collide two counter rotating beams of protons or heavy ions, with a tremendous energy of 7TeV. The beams move around the LHC ring inside a continuous vacuum guided by superconducting magnets maintained by a huge cryogenic system. To pinpoint the smallest fragments of the universe, Physicists have to build the biggest machine in the world!

The second photograph shows LIGO (Laser InterferometerGravitational Wave Observatory) another huge facility at Caltech where Gravitational waves, the weakest waves in the Universe have been successfully detected. The length of each of the arms in LIGO is 4 km and consequently, the sensitivity of the huge LIGO is 1 part in 10^{21} ! LIGO is basically an advanced Michelson's interferometer, whose arms are considerably long..4 kms

Preface

Physics: the word derives itself from the Greek language. Usually translated into English as “nature”, this ‘natural philosophy’, started dominating Science and Technology from its beginning itself. It seems that, all the disciplines in science and technology are linked with Physics. A few examples, which support this argument, are, Chemical Physics, Physical Chemistry, Geophysics, Biophysics, Astrophysics, Mathematical Physics and off course Engineering Physics. While the contributions of Newton, Gibbs, Carnot, Bernoulli, Young, Maxwell, Hertz, Gauss, Kirchhoff, Tesla, Joule, Ampere, Archimedes, Edison, and Faraday in the classical era remain unforgettable, the contributions of A few Nobel Laureates in the twentieth and the current century are also momentous. The first Nobel Prize bestowed on Roentgen in 1901 and the recent one on Andre Gim, and Nokolosov, for the discovery Graphene, are just a few examples. In short, the contributions of the Physics in technology leave no doubt for an enlightened. This books aims at providing the elements of Physics which are required to engineers.

*Dr. Narendra Mathakari,
narendra.mathakari@mitpune.edu.in
nlmathakari@gmail.com*

Before beginning, you may learn a few definitions

Science

Systemized knowledge derived through experimentation, observation, and study. Also, the methodology to acquire the knowledge

Technology

Definition 1: The application of *scientific advances* to benefit the humanity

Definition 2: The practical application of *science* to commerce or industry

Engineering

Definition 1: The design, construction and operation of structures and machines, using *scientific principles*.

Definition 2: Science, discipline, art and profession of acquiring and applying technical, *scientific* and mathematical knowledge to design and implement materials, structures, machines, devices, systems, and processes that safely realized desired objective or inventions.

Definition 3: The profession in which a knowledge of the mathematical and *natural sciences* gained by study, experience, and practice is applied with judgment to develop ways to use economically the materials and forces of nature for the benefit of mankind

Contents

CHAPTER 1

Interference

A basis of high precision instruments called Interferometers

CHAPTER 2

Diffraction

Diffraction plays a major role in high resolution optical instruments

CHAPTER 3

Polarization

Restricting the vibrations of light leads to applications

CHAPTER 4

Sound Engineering-I (Architectural Acoustics)

How to design enclosures with superior acoustics

Sound Engineering-II (Ultrasound)

The ‘soundless sound’ finds several applications in the technology

CHAPTER 5

Lasers

A coherent and amplified light; basis of photonics

CHAPTER 6

Semiconductor Physics

Fermi level determines the behavior of semiconductors and devices

CHAPTER 7

Basics of Quantum Mechanics (Wave particle duality)

Energy and matter are quantized

CHAPTER 8

A few applications of Quantum Mechanics (Wave-function& wave-equation)

Probing into the Physics of atoms, molecules and photons

CHAPTER 9

Superconductivity

Zero electrical resistance, but not yet at room temperature

CHAPTER 10

Elements of Nanotechnology

Tiny aggregates of atoms show new horizons to the technology

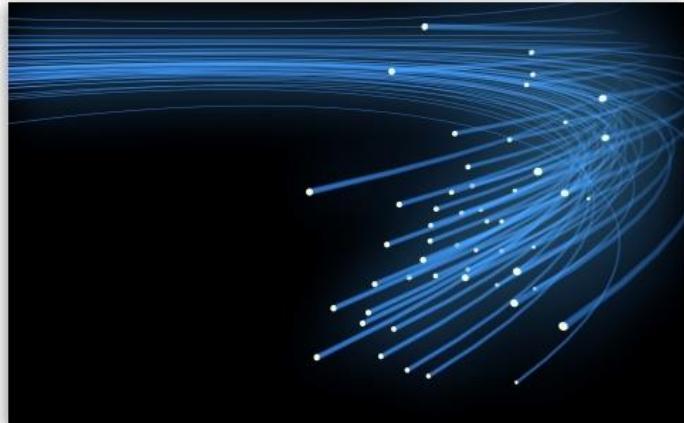
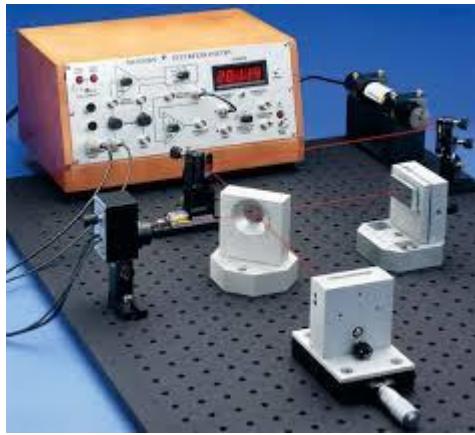


Albert Einstein (1879-955): Underestimated at school level, and then working as a clerk in the Swiss patent office, he at the age of 26, published his first three seminal papers, interpreting Brownian motion, relativity and the concept of quanta initiated by Max Planck. His contributions are too many to accommodate here, a few of which are $E = mc^2$, an acceptable interpretation of gravity as warping of space-time, concept of stimulated emission etc. He won Nobel prize in Physics for the quantum explanation of the Photoelectric effect; however, the mastermind deserved more than one. He correctly interpreted Gravity in terms of warping of the space-time. One of his endeavors, the unification of the four fundamental forces in the Universe was only partially fulfilled. In the Engineering Physics course, a thorough discussion of Quantum Physics will follow, encompassing the applications such as electronic configuration of the atoms, LASER, TRANSISTOR, Scanning Tunneling Microscope, Josephson junctions, SQUIDS etc. But, it is interesting to note that though he was the originator of Quantum Physics, never accepted it until his death, Probably only once, the Nature deceived the Genius. In 1917, he declared, "*for rest of the life I would reflect on what light is*".

Optics

For the rest of my life I will reflect on what light is

Albert Einstein (1917)

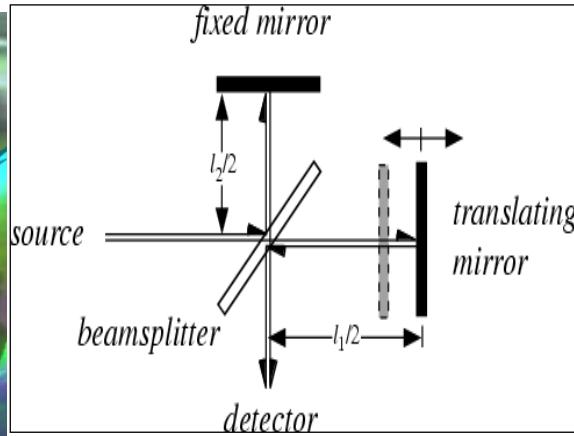


The photograph on the left shows one of the most versatile interferometers, the Michelson interferometer. There are several interferometers which form the basis of interferometry. The photograph on the right shows optical fibers which are essential components of modern communication systems. These are two out of several applications of Photonics, a rapidly emerging branch of Physics. The exact nature of light is yet to be understood. Thus, conventionally Optics, which describes light and its applications, divides itself into three branches, called Ray-Optics (reflection, total internal reflection, refraction, double refraction etc.), Wave-Optics (interference, diffraction and polarization) and Quantum-Optics (photoelectric effect, optoelectronics, lasers etc.).

Optics finds a few notable applications in day-to-day life as well as technology. A few of these are such as Optical Engineering, i.e. design and fabrication of several optical devices and instruments such as lenses, mirrors, gratings, prisms, interferometers, diffractometers, polarimeters, cameras, video cameras, movie projectors etc., Further, the disciplines appearing on the frontiers of current Physics, such as Photonics also require a thorough understanding of Optics.

CHAPTER 1

Interference



Interestingly, the oily films spread on the roads in rainy days, or the soup bubbles show beautiful colors. The pattern changes when viewed at different angles. This is due to thin film interference, where the thin film behaves like a ‘natural interferometer’. An extension of this concept gives birth to several interferometers, used for high precision measurements. What is an interferometer and what are its uses?

The answer to this question is in this chapter

Index

1.1 INTRODUCTION

How to produce a steady state and sharp interference pattern

1.2 THIN FILM INTERFERENCE

How thin film behaves as a natural interferometer

1.3 NEWTON'S RINGS (NEWTON'S INTERFEROMETER)

Circular fringes due to circular symmetry

1.4 MICHELSON'S INTERFEROMETER (Optional)

Circular fringes, but where is the circular symmetry?

1.5 INTERFERENCE COATINGS

How to minimize the unwanted reflection or transmission of light



Thomas Young (1773-1829): The Physicist with versatile intelligence was the first to demonstrate interference. He learned to read at the age of 2 and said to have read the complete Bible, twice, at the age of 6. He was a Physicist of high caliber as well as a Physician. He made significant contributions in Physics as well as Physiology, a few of which include understanding of the Physiology of Human eye, Physics of color vision, concept of modulus of elasticity known due to his name etc. He also gave an experimental foundation to the wave theory of light, which was based on his double slit experiment. What follows is a rapid revision of his experiment, necessary before going ahead.

1.1 INTRODUCTION

How to produce steady state and sharp interference pattern

The first demonstration of the interference occurred in 1801, due to Thomas Young, who, with the help of his famous double slit experiment, produced a well-defined interference pattern. He also proposed a theory of interference. His explanation was based on an assumption that light is a wave. Indeed all the phenomena related to interference conclusively prove that light is a wave.

In day to day life interference occurs in many situations. An oil film spread on the road in rainy days appears colored. This is due to interference. Interference is also responsible for the colored appearance of soap bubbles when illuminated by white light. Interference also has some interesting practical applications. It is the basis of high precision interferometers which are used for various measurements. Antireflection and anti-transmission coatings and interference filters are also based on interference. Interference is also used to inspect the preciseness of the optical devices such as glass plates, lenses etc.

Interference is the modification in the intensity when two or more waves are superimposed. It occurs only at a point where the waves overlap. As the waves surpass the point, they travel independently and unaffected.

The principle of superposition is given by

$$y = y_1 \pm y_2 \quad \dots(1.1)$$

Where y_1 and y_2 are the instantaneous displacements of first and second wave at the point of superposition. y is the resultant displacement. The \pm sign indicates that constructive as well as destructive interference are possible.

Let

$$y_1 = a_1 \sin wt \text{ and } y_2 = a_2 \sin(wt + \phi)$$

Where a_1 and a_2 are the amplitudes, w is the frequency and ϕ is the phase difference between the first and second wave.

By using the principle of superposition, it can be shown that

$$I = R^2 = a_1^2 + a_2^2 + a_1 a_2 \cos\phi \quad \dots(1.2)$$

Where R is the resultant amplitude and I is the resultant intensity. Eq. (1.2) can be analyzed with two special cases

If the phase difference $\phi = 0, \pm 2\pi, \pm 4\pi, \pm 6\pi, \dots \dots \pm 2n\pi$, then $\cos\phi = 1$. Then we have

$$I = R^2 = a_1^2 + a_2^2 + 2a_1 a_2 \cos(2n\pi) \quad \dots(1.3)$$

$$\Rightarrow I = (a_1 + a_2)^2 > a_1^2 + a_2^2$$

$$\Rightarrow I > I_1 + I_2$$

$$\text{if } a_1 = a_2 = a \text{ then } I = 4a^2$$

This indicates that if the phase difference between two waves is even multiple of π , then the resultant intensity is greater than the sum of the individual intensities. Thus intensity is enhanced. Such interference is called as constructive interference. The fringe with enhanced intensity is called maxima.

We know that the phase π corresponds to path $\frac{\lambda}{2}$. Thus the condition of constructive interference in terms of λ is

$$\text{Path difference for constructive interference} = 2n\frac{\lambda}{2} = n\lambda$$

If the phase difference $\phi = \pi, 3\pi, 5\pi, \dots, (2n \pm 1)\pi$ then $\cos\phi = -1$. Then we have

$$\begin{aligned} I = R^2 &= a_1^2 + a_2^2 + 2a_1 a_2 \cos(2n \pm 1)\pi \\ \Rightarrow I &= a_1^2 + a_2^2 - 2a_1 a_2 \\ \Rightarrow I &= (a_1 - a_2)^2 < a_1^2 + a_2^2 \\ \Rightarrow I &< I_1 + I_2 \end{aligned}$$

$$\text{If } a_1 = a_2 = a, \text{ then } I = 0$$

Thus if the phase difference between the two waves is $(2n \pm 1)\pi$, then the resultant intensity is less than the sum of individual intensities. If the amplitudes of the two waves are equal then the resultant intensity is zero. Thus the intensity is decreased. Such interference is called as destructive interference. The fringe with decreased intensity is called as minimum. The condition of the destructive interference in terms of λ is

$$\text{Path difference for destructive interference} = (2n \pm 1) \frac{\lambda}{2}$$

Figure (1.1) shows the interference pattern with maxima and minima

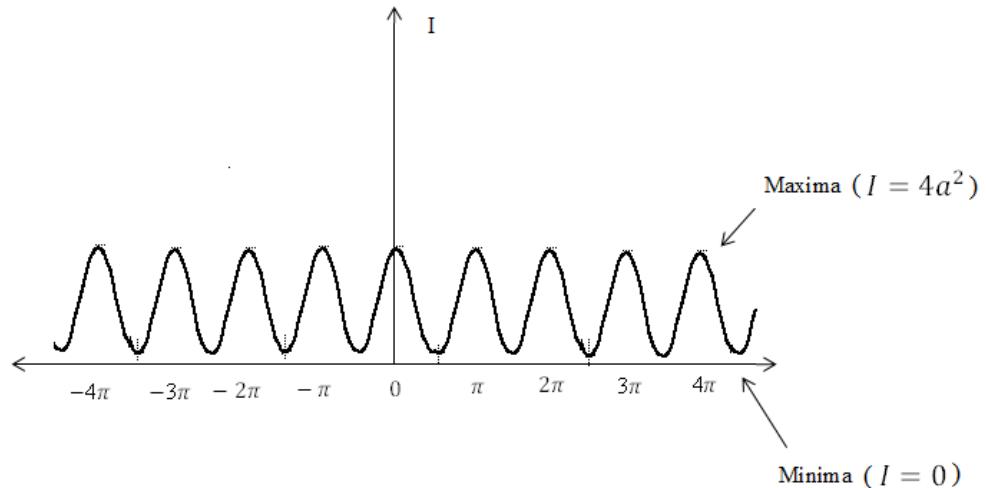


Figure (1.1): A typical interference pattern

Apart from the conditions of maxima and minima, other path differences are also possible. Thus depending upon the path difference, the resultant intensity may take any value from 0 to $4a^2$. It is to be noted that, in interference, the energy is neither created nor destroyed. At maximum there is an enhancement of intensity. But this is not the created energy. In minima there is decreased intensity. But this is not the destroyed energy. Indeed, in interference, there occurs a redistribution of the energy. The intensity which was supposed to be at minimum is transferred to maximum. Thus interference can also be defined as the redistribution of energy due to superposition of the waves.

Fig 1.1 shows a well-defined interference pattern consisting of maxima and minima. For obtaining a well-defined, sharp and steady state interference pattern, some conditions have to be satisfied. These are

- i. The waves must be coherent. This means that the phase difference between the interfering waves at any point should remain constant with respect to time. If there is no coherence, the phase difference and hence the intensity will change with respect to time. The pattern will not be a steady state pattern. Coherence is the prime condition to be satisfied for getting a steady state interference pattern
- ii. For having a sharp, distinct and well defined interference pattern, the waves must be monochromatic and they should have same amplitude. If these conditions are not satisfied then the intensity at maxima will not be maximum and the intensity at the minima will not be zero. Thus a distinct and sharp interference pattern cannot be obtained

- iii. There must be a path difference.
- iv. There should be a systematic and gradual variation of the path difference. For having zeroth, first, second, third maxima etc., the path difference should gradually increase from 0, λ , 2λ , 3λ Similarly, for having the first, second, third minima at appropriate places, the path difference at these points should take values from $\frac{\lambda}{2}$, $\frac{3\lambda}{2}$, $\frac{5\lambda}{2}$ etc. Thus for having alternate maxima and minima, there needs to be a gradual and systematic variation of the path difference.
- v. The path difference should vary only with the position and not with time.

Two separate and independent ordinary sources of light are always incoherent. If a screen is illuminated by such sources, it does not show a steady state interference pattern, as the phase difference between the waves at any point keeps on changing rapidly and randomly. Thus intensity changes several times in a second. Therefore the screen appears to be uniformly illuminated.

The prime condition for interference is coherence. Two independent sources of light can be monochromatic but not coherent (laser is an exception). Coherence is possible when the two sources are derived from the same source. This can be done by using two techniques, namely division of wave-front and division of amplitude. Fig.1.2 shows the technique of division of wavefront. A point monochromatic source of light emits primary (spherical) wavefront. Two point like slits become two sources of light, which emit two secondary (spherical) wavefronts. Point sources (spherical wavefronts) are employed in this technique. Both these wavefronts are coherent as they are derived from the same source. The technique of division of the wavefront is used in Young's double slit experiment

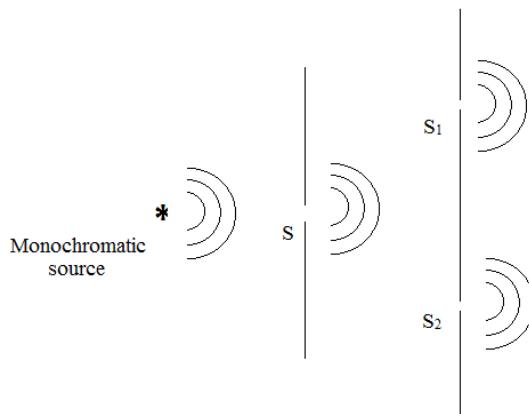


Figure 1.2Division of wavefront

Fig.(1.3) depicts the technique of division of the amplitude (intensity). A semi-silvered glass plate is inclined at 45° . A ray of light having amplitude A is incident on this glass plate. As the glass plate is semi-silvered, part of this ray is transmitted and part is reflected. The intensities of both these rays are approximately equal to $\frac{I}{2}$. This technique requires extended source (parallel wavefront). The division of amplitude technique is used in thin film interference, Newton's

ringsexperiment and Michelson's interferometer.

Apart from these two techniques, refraction and reflection are also used to obtain coherent sources. Fresnel's biprism and Lloyds mirror are based on these principles.

In this chapter we will learn the interference based on thin films and its applications

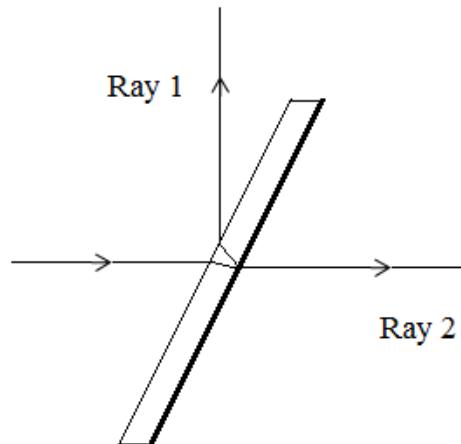


Fig 1.3 Division of amplitude

If all the conditions required for generating steady state and sharp interference pattern are satisfied then a well-defined and steady state pattern can be produced. The arrangement which takes care of all these conditions and which generates a steady state and sharp interference pattern is called as interferometer. There are varieties of interferometers used for various applications. Indeed the Young's double slit apparatus can be called as 'Young's interferometer'. In the subsequent sections, we will see that thin film behaves as a natural interferometer. Newton's rings apparatus may be treated as Newton's interferometer. The other examples are Michelson's interferometer, Fabry Perot interferometer etc. The Newton's interferometer and Michelson's interferometer are based on the thin film interference.

1.2 THIN FILM INTERFERENCE(**Compulsory but derivation is optional**)

How thin film behaves as a natural interferometer

Thin film is a transparent film having thickness slightly above the average wavelength of light. The average wavelength of visible light is roughly 5500 \AA° ($0.55 \mu\text{m}$). Thus in optics, a film having a thickness in mm is also considered as a thick film. There are several examples of thin films in the real world. These are oil films on the roads in rainy days, soap bubbles etc. Thin films are encountered in technology also. The antireflection coatings on camera lenses and solar cells, anti-transmission coatings on invisible glasses, interference filters etc are based on thin film interference. When exposed to light, the thin film produces an interference pattern. The thin film interference was first observed by Newton and Robert Hook. It was Thomas who gave the correct explanation to the phenomenon.

Fig 1.4 shows the ray diagram of the thin film interference. A thin film having uniform thickness t , refractive index μ is exposed to a monochromatic ray of light from an extended source having wavelength λ . The media above and below the film have refractive indices lower than that of film. (Freely suspended soap bubble in air is an example of this case). This is one out of a few more cases which will be discussed later.

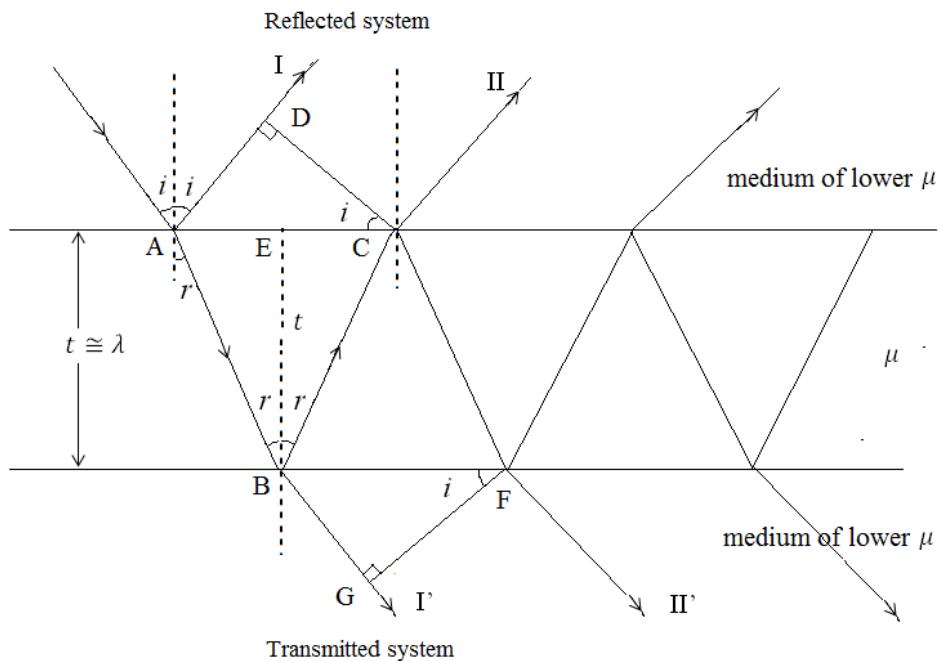


Figure 1.4 Thin film interference; the diagram is idealized and represents one out of a few more cases

A monochromatic ray from an extended source is incident on the film at the point A, at an angle of I and is reflected at i itself. During refraction, it enters into a denser medium at an angle of r . On the second interface at point B, the ray is incident at r and is reflected at r itself. Such reflections and refractions continue within and outside the film as shown in the Fig 1.. During every reflection and the refraction, the light suffers a loss in the intensity and thus the amplitude of various rays in the reflected and the transmitted system decreases gradually. Thus, here the conditions required for generating sharp fringes are not exactly satisfied. However, ray I and II (and consecutive rays) in the reflected as well as transmitted system are coherent as these are derived from the single ray. (This is amplitude division technique). Ray I and II, though appear parallel, can still interfere if watched through the lens in the eye. Or otherwise, if observed through the microscope, the lens will make them interfere. Now the interference of ray I and II will decide the intensity of the point A. Note that the rays next to the ray II are weak in intensity and therefore can be ignored. Point A will appear bright if I and II interfere constructively and it will appear dark if destructive interference takes place. As we know, the nature of the interference is essentially governed by the path difference between the interfering rays.

Ray II, before it arrives at C, travels a distance of $AB + BC$, while with respect to the line CD, ray I travels a distance of only AD . Additionally, as the distance traveled by the ray II is through the denser medium, it becomes slower and spends more time in the medium depending upon μ . Thus the geometrical path ($AB + BC$) should be converted into an optical path $\mu(AB + BC)$. In other words, ray II suffers a decrease in the wavelength as it travels through the denser medium. The decrease is governed by μ . Thus the optical path of ray II with respect to the line CD is larger than ray I. (*refer appendix I to chapter 1*). Thus we get

This Derivation is Optional

$$PD_{R,I,II} = \mu(AB + BC) - AD \quad \dots(1.4)$$

For the generalization of the Eq(1.4), we need to express PD in terms of t , μ and i (or r). The geometry of the Fig 1.4 assists in doing the task

Ideally, t and μ are same at all the points; thus the symmetry suggests,

$$AB = BC \Rightarrow AB + BC = 2AB$$

From the geometry of the triangle ABE, we have

$$AB = \frac{t}{\cos r} \quad \dots(1.5)$$

From the geometry of the triangle ADC, we have

$$\begin{aligned} AD &= AC \sin i \\ AD &= (AE + EC) \sin i \\ AD &= 2AE \sin i \end{aligned}$$

From triangle AEB,

$$AE = \tan r \times t$$

Substituting AE, we get

$$AD = 2\tan r \times t \times \sin i$$

Using Snell's law

$$AD = 2\tan r \times t \times \mu \sin r \quad \dots(1.6)$$

Substituting AD and AB from eqn 1.5 and 1.6 into 1.4, we have

$$PD_{R,I,II} = \mu\left(\frac{2t}{\cos r}\right) - 2t \times \tan r \times t \times \mu \sin r$$

$$PD_{R,I,II} = \left(\frac{2\mu t}{\cos r}\right) - \left(\frac{2\mu t}{\cos r}\right) \sin^2 r$$

$$PD_{R,I,II} = \frac{2\mu t}{\cos r} (1 - \sin^2 r)$$

$$PD_{R,I,II} = \frac{2\mu t}{\cos r} (\cos^2 r)$$

Read from here (Compulsory)

$$PD_{R,I,II} = 2\mu t \cos r \quad \dots(1.7)$$

As the ray I is reflected from the denser medium, according to Stokes law, it's phase is reversed by $= 180^\circ$ (*that is by $\frac{\lambda}{2}$*). However, this is not the case with ray II, as it is transmitted at A, reflected at the point B, but due to the rarer medium, and then transmitted at C. (According to Stoke's law, there is no phase reversal during transmission, or reflection, but due to rarer medium. *(For a detail account of Stokes law, refer Appendix II to chapter 1)*). As ray I is reversed in phase and ray II is not, in this case, there is an additional path difference of $\pm \frac{\lambda}{2}$ between ray I and II.

Thus,

$$PD_{R,I,II} = 2\mu t \cos r \pm \frac{\lambda}{2} \quad \dots(1.8)$$

Eq. (1.8) accounts for the path difference between consecutive reflected rays. Note that the presence or absence of the factor $\pm \frac{\lambda}{2}$ in such equation essentially depends upon the situation i.e. the relative denseness of film w.r.t. to the media above and below it.

We now use the Eqns(1.7) and (1.8) to identify some interesting characteristics of thin film interference.

- i. *The interference patterns of the reflected side and transmitted side of the thin film are always complimentary*

As t and μ are assumed to be same throughout, the geometry of the triangles ABC and BCF and triangles ACD and BFG is same. Thus the geometrical path difference between reflected rays I and II and transmitted rays I' and II' is same. However, the interference patterns i.e. occurrence

of maxima (or minima) from the reflected side and the transmitted side must complement each other, otherwise the optical energy will not be conserved. Thus for transmitted rays we have

$$PD_{I',II'} = 2\mu t \cos r \quad \dots(1.9)$$

Note that if Stokes law is carefully applied for the reflected rays and transmitted rays, then if $\pm \frac{\lambda}{2}$ is present in the equation for the path difference for reflected rays, and then it will be obviously absent for transmitted rays and vice versa. If a point appears bright at a given angle in the reflected system, then it appears dark at the same angle in the transmitted system and vice versa. Note that in the transmitted system, ray I' does not undergo any phase reversal, as it is just transmitted at point A and B. Ray II' is reflected at B and C and transmitted at F. As the reflection at B and C is due to rarer medium, there is no phase reversal. Thus both I' and II' do not undergo phase reversal and therefore the term $\pm \frac{\lambda}{2}$ is absent in the transmitted system. The nature of interference (constructive or destructive) between I and II or I' and II' decides the intensity at the point A (bright or dark). Similar discussion is possible for the rays incident at the points, other than A.

In the present case (when the medium above and below the film is rarer than the film) for reflected system

For constructive interference

$$2\mu t \cos r \pm \frac{\lambda}{2} = 2n \frac{\lambda}{2}$$

For destructive interference

$$2\mu t \cos r \pm \frac{\lambda}{2} = (2n \pm 1) \frac{\lambda}{2}$$

Note that, depending upon the situation, that is relative denseness of the film with respect to the medium on the top or bottom, the factor $\pm \frac{\lambda}{2}$ may or may not be present on LHS

ii. An extremely thin or extremely thick film cannot produce interference pattern

For, ex. if the film is too thin ($t \ll \lambda$) then $t \rightarrow 0$, thus the P. D., irrespective of any point on the film has same path difference, say $\pm \frac{\lambda}{2}$ (or \sim zero). Thus all points appear dark in the reflected system and bright in the transmitted system (or vice versa). However, this is not the interference pattern, as it requires the presence of alternate maxima and minima. In case of the films having their thickness too large as compared to wavelength of light also the interference pattern cannot be observed. Owing to very large value of t , the P.D. at every point becomes very large as compared to wavelength of any color in the light. Thus, for every color, there exist a few integers

for which the P.D. is an odd multiple of $\frac{\lambda}{2}$ and there also exist a few integers for which the P.D.

will be an even multiple of $\frac{\lambda}{2}$. Thus, at every point the constructive as well as destructive interference of any color is possible. Thus interference pattern will not be observed.

iii. *Fizau's and Haidinger's fringes:*

If we expect the thin film to produce an interference pattern, then it should be able to produce alternate maxima and minima. Thus the path difference should gradually pass through the even and odd multiples of $\frac{\lambda}{2}$. As, it can be observed, it is not possible in the case depicted in the Fig 1.4. Here every parameter including t , μ and r remains the same at all the points. The systematic and gradual variation of path difference requires that either t should vary by keeping r same or vice versa. Both these cases are depicted in the Fig 1.5a and b.

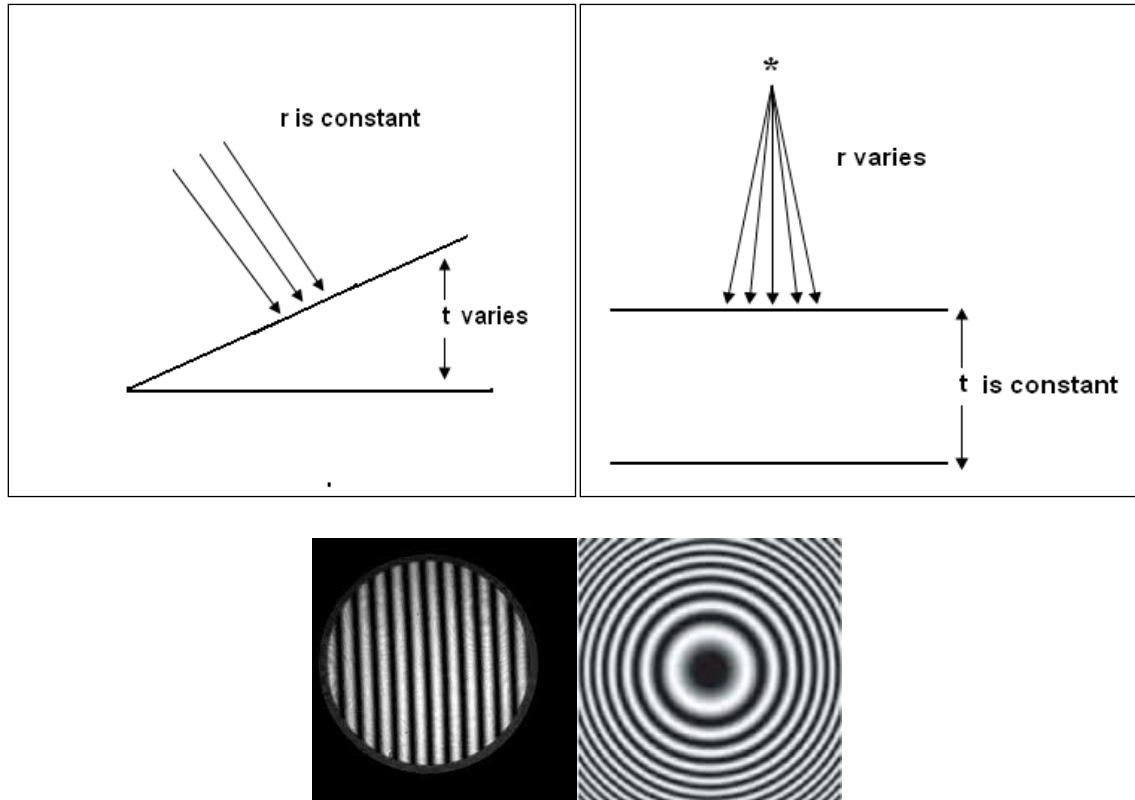


Figure 1.5: Fizau's and Haidinger's fringes

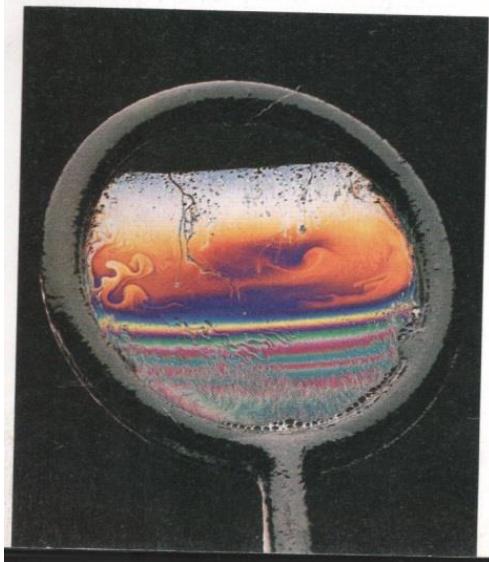
As it can be observed, in the first case, t increases gradually and, as the wavefront is parallel, r remains the same. As the variation of the t occurs in horizontal (X) direction, the P.D. and the change in the intensity of the fringe occurs in horizontal (X) direction. However,

*t*remainssame along a direction parallel to edge of the film and perpendicular to the plane of the paper. Thus P.D and consequently the intensity of the fringe remain same along the edge of the film(Z). These fringes, which are parallel to the edge of the film, equidistant, and in the horizontal plane, are referred as **Fizau's fringes**.

In Fig 1.5.b, as the source is a point source, it emits a spherical wavefront and thus the rays are incident on the film along various cones. Thus, for each cone, r remains constant over a circle. Thus PD remains constant over a circle. Owing to this circular symmetry, if observed from the top, the fringes will appear concentric and circular. This are referred as **Haidenger's fringes**. Note that in Fizau's fringes t varies and r remains the same, while in Haidenger's fringes r varies and t remains the same. We can think about a situation, where t as well r is varied, but then interference pattern will be complicated

iv. *Why oil films on wet roads and soap bubbles appear colored:*

Eqn. 1.8 also explains why the oil films spread on the road in rainy days or the soap bubbles appear colored. Such films are exposed to the ambient light, which is polychromatic. Further the parameters like t , r and μ vary randomly. Thus the P.D. varies from region to region in a random manner. Such P.D. satisfies the conditions of constructive and/or destructive interference for different colors in the different regions in a random manner. Thus some colors are enhanced in some regions and some are suppressed in the other regions. We thus see a random, but beautiful distribution of colors.



Colors in the soap film are due to thin film interference. Why there are no colors near the upper edge?



Wings of the Morpho butterfly. The colors are due to thin fil interference. The colors change when viewed at different angles. Why?

Interestingly, the appearance changes with the angle of viewing (variation of r). It is also needless to mention that the pattern of colors observed in the reflected system will exactly complement the pattern in the transmitted system.

Example (1.1): A thin film spread on a road which is optically denser than the film. Thickness of the film is $0.55 \mu\text{m}$, while its refractive index is 1.34. Why does the film appear greenish, when viewed in the reflected mode at 68° ? The wavelength of the green light is $\sim 5500 \text{ Å}^\circ$

Solution:

As the film is denser than air and as road is denser than the film, the ‘Stoke’s phase reversal occurs for both the rays, and the factor $\pm \frac{\lambda}{2}$ disappears from the eqn for P.D.. Thus, the equation for the P.D. for reflected rays is

$$P.D_{R, I, II} = 2\mu t \cos r$$

The film will appear greenish, if constructive interference occurs for the green color

Thus,

$$2\mu t \cos r = n\lambda$$

As the film is thin, considering least possible value of $n (=1)$

$$2\mu t \cos r = \lambda$$

Angle of viewing is same as angle of incidence, thus

$$2\mu t \cos i = \lambda$$

$$2 \times 1.34 \times 5500 \cos i = 5500$$

$$i = 68^\circ$$

Thus constructive interference occurs in this case when the angle of viewing is 68° . The film thus appears greenish, when viewed at 68° . The film, at the same angle but from the opposite side will appear less greenish i.e. more purplish.

Example 1.2: A thin film of CCl_4 having refractive index 1.46 and thickness $0.1068 \mu\text{m}$ is spread on water having refractive index 1.33. If viewed at 45° which color will be seen enhanced?

Solution:

Angle of viewing = angle of incidence

$$\mu = \frac{\sin i}{\sin r}$$

$$\Rightarrow 1.46 = \frac{\sin 45}{\sin r}$$

$$\Rightarrow \sin r = \frac{\sin 45}{1.46}$$

$$\Rightarrow \sin r = \frac{\sin 45}{1.46}$$

$$\Rightarrow \sin r = \frac{\sin 45}{1.46}$$

$$\Rightarrow \sin r = 0.48$$

$$\Rightarrow r = 28.97^\circ$$

We have

$$2\mu t \cos r \pm \frac{\lambda}{2} = 2n \frac{\lambda}{2} \text{ (Constructive interference)}$$

(Note that as the substrate (water) is rarer than the film, and as the medium above the film is air, the Stoke's factor $\pm \frac{\lambda}{2}$ is present on the LHS)

$$\Rightarrow 2\mu t \cos r = (2n \pm 1) \frac{\lambda}{2}$$

As the film is thin, we will choose the smallest value of n for which the RHS will be nonzero. The n = 0. Subsequently, we will prove that n = 1 is not possible

$$\Rightarrow 2\mu t \cos r = \frac{\lambda}{2}$$

$$\Rightarrow 2 \times 1.46 \times 1068 \times \cos 28.97 = \frac{\lambda}{2}$$

$$\lambda = 5456 \text{ } \text{\AA}^\circ \Rightarrow \text{Green color will be enhanced}$$

Let us choose n = 1

$$2\mu t \cos r = (2n \pm 1) \frac{\lambda}{2}$$

$$\Rightarrow 2 \times 1.46 \times 1068 \times \cos 28.97 = \frac{3}{2} \lambda$$

$$\Rightarrow \lambda = 1818.89 \text{ } \textit{\AA}$$

This wavelength is beyond the visible spectrum, therefore it is necessary choose $n=0$. Here
 $n \geq 1$ has no significance.

1.2 WEDGE SHAPED FILMS(**Compulsory, but derivation is optional**)

Fringe width depends upon the wedge angle

Following Ray Diagram is compulsory



Refer the ray diagram as shown in Figure 1.6

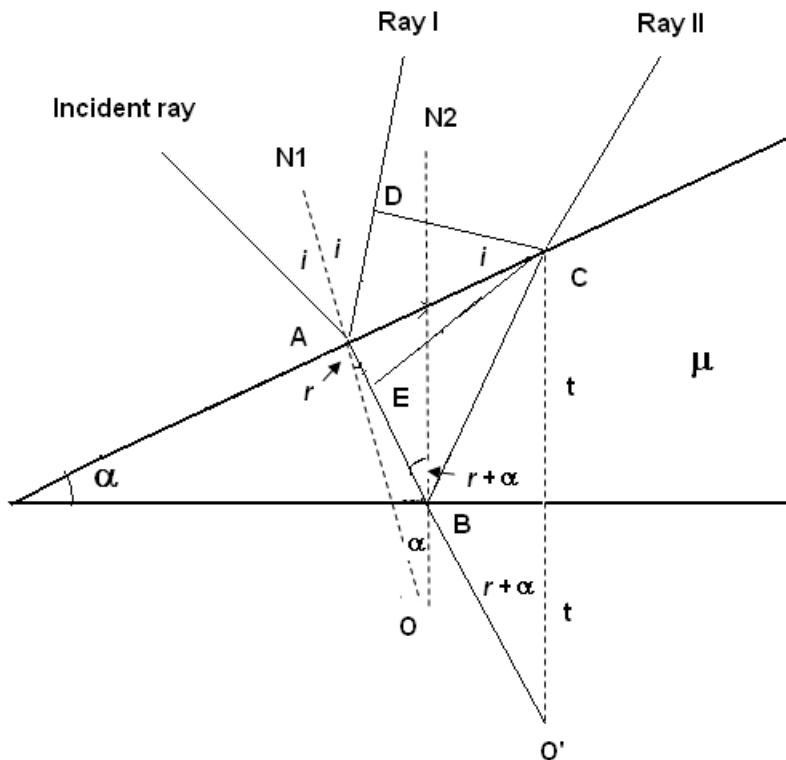


Figure 1.6 Ray diagram for the Wedge shaped film

Fringe width

The P.D. between the reflected ray I and II can be shown to be equal to

$$P.D_{W,R,I,II} = 2\mu t \cos(r + \alpha) \pm \frac{\lambda}{2} \quad \dots(1.10)$$

(This formula is derived in Appendix III to chapter 1)

Optional 

For normal incidence $r \rightarrow 0$. Thus eqn (1.10) simplifies to

$$P.D_{W,R,I,II} = 2\mu t \cos \alpha \pm \frac{\lambda}{2} \quad \dots(1.11)$$

This path difference will produce n^{th} bright fringe, if t_n satisfies the following Eqn.

$$P.D_{W,R,I,II} = 2\mu t_n \cos \alpha \pm \frac{\lambda}{2} = n\lambda \quad \dots(1.12)$$

Similarly for the $(n+1)^{\text{th}}$ bright fringe

$$P.D_{W,R,I,II} = 2\mu t_{(n+1)} \cos \alpha \pm \frac{\lambda}{2} = (n+1)\lambda \quad \dots(1.13)$$

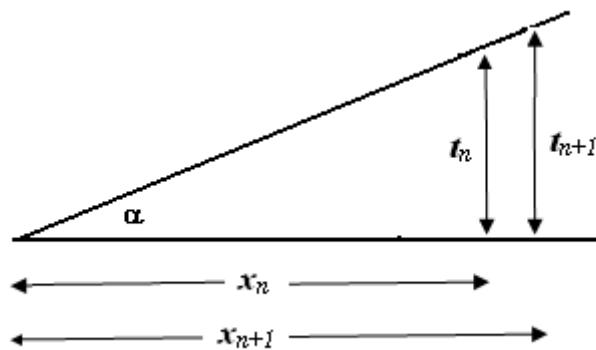


Figure 1.7 Fringe width

Eqn (1.13) shows that at $t = 0$, 1st dark fringe will be produced. Thus, in between nth and (n+1)th bright fringe, (n+1)th dark fringe will appear. Consider Fig. 1.7. We observe that

$$t_n = x_n \tan \alpha \text{ and } t_{n+1} = x_{n+1} \tan \alpha$$

Substituting both these identities in eqn 1.12 and 1.13, we get

$$P.D_{W,R,I,II} = 2\mu x_n \tan \alpha \cos \alpha \pm \frac{\lambda}{2} = n\lambda \quad \dots(1.14)$$

$$P.D_{W,R,I,II} = 2\mu x_{(n+1)} \tan \alpha \cos \alpha \pm \frac{\lambda}{2} = (n+1)\lambda \quad \dots(1.15)$$

Subtracting eqn (1.14) from (1.1.15) and rearranging, we get,

Compulsory ↓

$$\boxed{x_{n+1} - x_n = \frac{\lambda}{2\mu \tan \alpha \cos \alpha} = \frac{\lambda}{2\mu \sin \alpha} \approx \frac{\lambda}{2\alpha} (\text{air})} \quad \dots(1.16)$$

Above formula represents the width of a dark fringe. Similar procedure for the bright fringe will yield the same result. Note that this formula represents only approximate fringe width, as it has been obtained with a few assumptions.

Example 1.3: As shown in the Fig. 1.6, a wedge is formed by separating two glass plates by an extremely thin wire kept at a distance of 10.0 cm from the edge. When illuminated by sodium light of wavelength 5890 Å° the width of the Fizeau's fringes is measured to be 2.945 mm. Calculate the diameter of the wire

Solution:

$$fw = \frac{\lambda}{2\mu \tan \alpha}$$

$$2.945 \times 10^{-3} = \frac{5890 \times 10^{-10}}{2 \times 1 \times \tan \alpha}$$

$$\tan \alpha = 1 \times 10^{-4}$$

$$\alpha = 5.7296 \times 10^{-3} \text{ deg}$$

From the figure

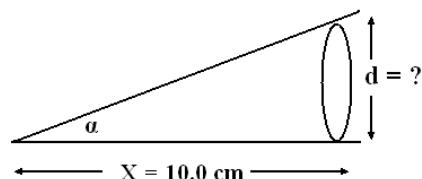


Figure 1.8 Wedge film for Ex 1.3

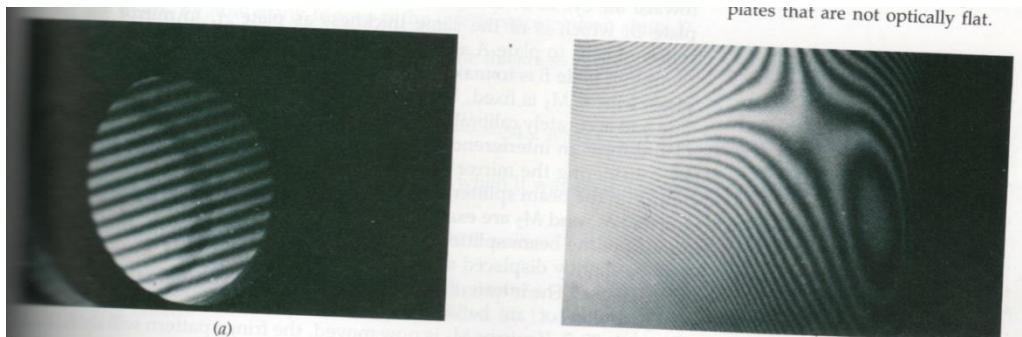
$$d = X \tan \alpha$$

$$d = 10.0 \times 1 \times 10^{-4} \text{ cm}$$

$$d = 1 \times 10^{-3} \text{ cm}$$

$$d = 10 \mu\text{m}$$

Thus, by measuring the fringe width, we can measure the extremely small dimensions such as diameters of thin wires. This method can also be extended to any thin spacer that can be used to form the wedge. From the Eqn. 1.16), one may note that thinner the dimensions of the object, more it is easy to measure it with the help of this technique. This is because the fringe width increases with the decrease in the dimensions of the object to be measured. On the contrary, the measurement by conventional techniques such as screw gauges and Vernier becomes more difficult as the object becomes thin. We thus conclude that optical measurements are more convenient and precise than the conventional measurements.



Fringes due to wedge shaped film (Regular and Irregular): The Optical flatness of a glass slab can be checked.



Sir Isaac Newton (1642-1727) While the fall of an apple remains an ordinary phenomenon for a layman, it was not so for him, as the phenomenon inspired him to discover the universal law of Gravitation. He also concluded that, the force which pulls apple down also makes Earth rotate around Sun and Moon around the Earth. While his several contributions such as mechanics, optics and gravitation are noteworthy, it is suffice to mention that he is also claimed to be an originator of calculus. He acquired the Lucasian Professorship of Mathematics in Cambridge University, just at the age of 26, and was elected as a Fellow of Royal Society at the age 30. In his famous book named PRINCIPIA, he published the three laws of motion and the Universal law of Gravitation. In another publication named OPTICKS, he presented Physics of the spectrum, interference, color vision and rainbow. Herewith, we discuss the circular and concentric interference fringes known by his name, as he was the first to observe.

Compulsory, but derivations are optional



1.3 NEWTON'S RINGS (NEWTON'S INTERFEROMETER):

Circular fringes due to circular symmetry

The thin film may be treated as a ‘natural interferometer’ as it produces an interference pattern. Our previous discussion of the thin film interference is oversimplified, however, in nature, the cases of thin films can be quite complicated. However, at this stage it is possible to verify whether the concept of the thin film interference can be extended to design an interferometer. One such interferometer was ‘designed’ by Isaac Newton. He didn’t call his system as an interferometer, and as he proposed the corpuscular theory of light, neither did he understand the Physics behind the production of Newton’s rings. The phenomenon was correctly explained by Thomas Young using his wave theory of light.

Consider Fig. 1.6. The figure is exaggerated because, the planoconvex lenses involved in the practice have extremely small curvature, and the incident ray falls almost along the normal.

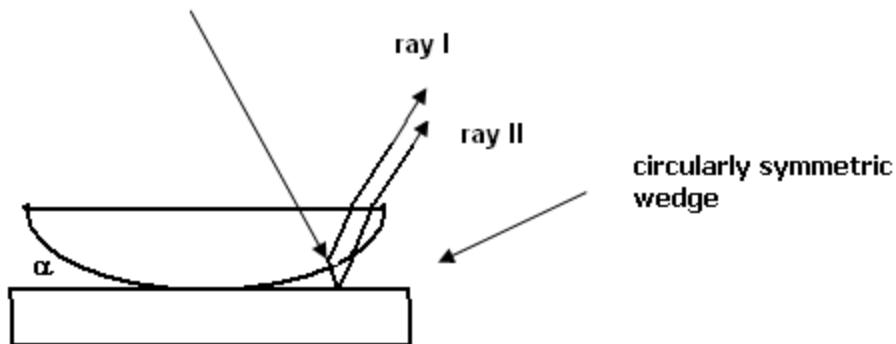


Figure 1.9The ray diagram for Newton's rings

It can be understood that, in this case the contours of constant thicknesses i.e. constant path difference form circles. Thus the Fizau’s fringes in this case are circular and concentric. The film involved here is a special case of wedge shaped films. Thus, we can proceed ahead with eqn (1.10)

$$P.D_{I,II} = 2\mu t \cos(r + \alpha) \pm \frac{\lambda}{2} \quad \dots(1.17)$$

Optional



Some assumptions such as $\alpha \rightarrow 0$, $r \rightarrow 0$ will simplify the discussion. Thus we have, for n^{th} dark ring.

$$P.D_{I,II} = 2\mu t_n \pm \frac{\lambda}{2} = (2n \pm 1) \frac{\lambda}{2}$$

We get,

$$P.D_{I,II} = 2\mu t_n = n\lambda \text{ (n}^{\text{th}} \text{ dark ring)} \quad \dots(1.18)$$

The geometry of the Figure 1.9 helps to express t_n in terms of parameters of interest such as R , the radius of curvature of the planoconvex lens and D_n , the diameter of the n^{th} dark ring. Figure 1.7 clearly indicates that

$$R^2 = (R - t_n)^2 + r_n^2$$

On solving the above eqn and by approximating $t_n^2 \rightarrow 0$, we get

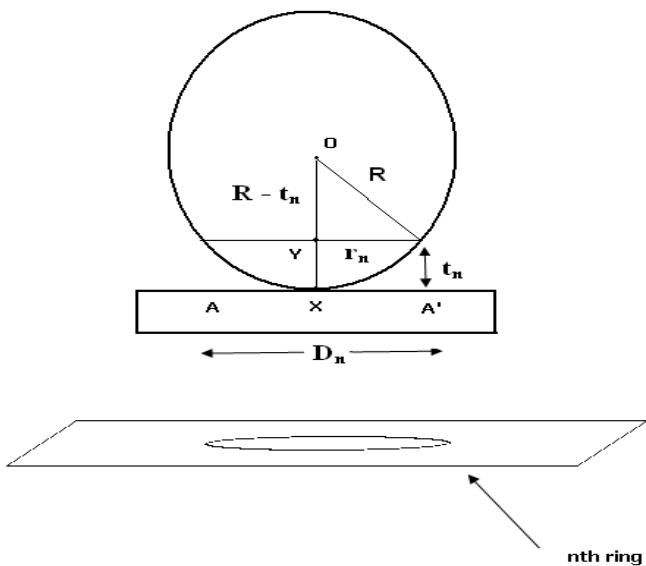


Figure 1.9 How t_n relates with R and D_n

$$t_n = \frac{D_n^2}{8R}$$

Substituting above t_n in the eqn (1.13), we get

$$2\mu \frac{D_n^2}{8R} = n\lambda \quad \dots(1.19)$$

Compulsory ↓

Thus,

$$D_n^2 = \frac{4Rn\lambda}{\mu} \quad \dots(1.20)$$

From above eqn , we get

$$D_n = \sqrt{\frac{4Rn\lambda}{\mu}} = \sqrt{\frac{2R\lambda}{\mu}} \sqrt{2n} \quad \dots(1.21)$$

$$\Rightarrow D_n \propto \sqrt{2n} \quad \dots(1.21)$$

A similar derivation for the m^{th} bright ring yields

$$D_m = \sqrt{\frac{2R\lambda}{\mu}} \sqrt{(2m \pm 1)} \quad \dots(1.22)$$

$$\Rightarrow D_m \propto \sqrt{(2m \pm 1)} \quad \dots(1.22)$$

From eqns 1.21 and 1.22, we may be tempted to conclude that, the diameters of dark rings are proportional to the square root of even natural numbers, and the diameters of bright rings are proportional to square root of odd natural numbers. However, this need not be the case always, as the Stoke's factor in the eqn 1.13 will disappear in a few situations, say if the μ increases gradually from planoconvex lens to the wedge to the optical flat or it may decrease also. On disappearance of the Stoke's factor, the equations 1.21 and 1.22 will complement each other. Thus, it is better to state that **diameters of Newton's rings are proportional to the square root of the natural numbers.**

The wedge angle in this case increases gradually, and thus the width of Newton's rings decreases with the sequence number of the ring. Further, it is difficult to keep the center of the fringes dark or bright permanently. Thus it is difficult to decide sequence number of the ring. Thus eqn 1.16 will not yield the same accuracy, and the accuracy will depend upon which ring is measured. Thus to have better accuracy, eqn 1.16 can be written for the m^{th} dark ring and then the difference, $(D_m^2 - D_n^2)$ can be obtained. Thus,

$$\boxed{D_m^2 = \frac{4Rm\lambda}{\mu}} \quad \dots(1.23)$$

Taking the difference of the eqn 1.23 and 1.20 and rearranging,

$$\boxed{R = \frac{\mu(D_m^2 - D_n^2)}{4(m-n)\lambda}} \quad \dots(1.24)$$

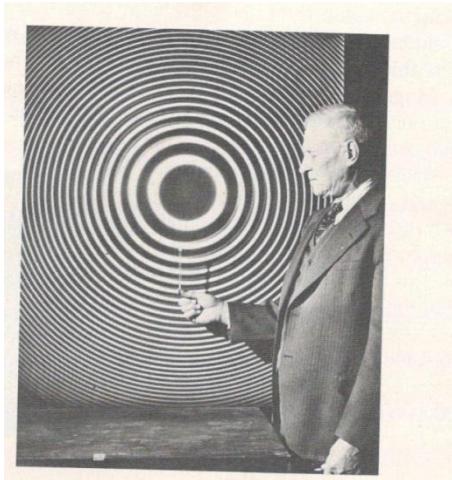
Eqn 1.21 suggests a few applications of Newton's interferometer. If the diameters D_m and D_n of the m^{th} and n^{th} dark rings are measured, and if μ is known, then the radius of curvature R of a planoconvex lens can be calculated provided λ is known. R can also be measured by spherometers, thus λ can also be calculated. Further, if R , and λ are known then μ can be evaluated.

Eqn. (1.20) suggests that if the medium having refractive index μ is replaced by the air, then the diameter of Newton's ring will increase. By choosing new symbols, we can thus write

$$D_n^2' = \frac{4Rn\lambda}{\mu} \text{ and } D_n^2 = 4Rn\lambda$$

Thus we have,

$$\boxed{\mu = \frac{D_n^2}{D_n^2'}} \quad \dots(1.25)$$



Newton's rings



Newton's rings due to a lens during the manufacturing Process. What is signified by the irregularity?

For better accuracy,

$$\mu = \frac{D_m^2 - D_n^2}{D_m^2 + D_n^2} \quad \dots(1.26)$$

Thus the Newton's interferometer also provides a method to measure the refractive index of liquids. It is needless to mention that accurate measurements require a monochromatic source.

Example 1.4 The diameters of the fifth and the third dark rings as measured by Newton's interferometer are 0.343 cm and 0.266 cm respectively. If the setup is operated by a sodium light of wavelength 5890 \AA , calculate the radius of the curvature of the Planoconvex lens. Assume that the setup is kept in the air.

We have

$$R = \frac{\mu(D_m^2 - D_n^2)}{4(m-n)\lambda}$$
$$R = \frac{1 \times (0.343^2 - 0.266^2)}{4 \times (5-3) \times 5890 \times 10^{-8}}$$
$$R = 99.52 \text{ cm}$$

Example (1.5): The diameters of the 5th and 3rd ring measured, when the Newton's ring set up is kept in a 30% sugar solution are 0.292 cm and 0.226 cm respectively. When the sugar solution is replaced by air, the diameters increase to 0.343 cm and 0.266 cm. respectively. What is the refractive index of the 30 % sugar solution?

We have

$$\mu = \frac{D_m^2 - D_n^2}{D_m^2 + D_n^2}$$
$$\mu = \frac{(0.343^2 - 0.266^2)}{(0.292^2 - 0.226^2)}$$
$$\mu = 1.37$$

Compulsory



Characteristics of Newton's rings

- i. Newton's rings on reflected side are complementary to the Newton's rings on transmitted side
- ii. If the glass plate in the Newton's ring set up is replaced by the Mirror, then Newton's rings fade out and a uniform illumination is observed.
- iii. If the Newton's ring set up is illuminated by white light then a few colored rings near the center are observed.
- iv. When there is air gap at the center, the ring at the center may appear bright. .
- v. If the lens is gradually lifted up, then the Newton's rings are shifted outwards
- vi. If the monochromatic source in the setup is replaced by a source of higher wavelength, then the diameters of Newton's rings are increased.
- vii. If the planoconvex lens in the setup is replaced by the planoconvex lens of higher radius of curvature then the diameters of the rings will increase



Albert A. Michelson (1852-1931) His specialty was high precision optical instruments. For several years, his measurement of the speed of the light was considered to be authentic. He redefined the meter in terms of the wavelength of light. Using his interferometer, he measured the diameters of the stars, which appear just like point objects on the Earth. One of his greatest achievements is confirming the absence of 'ether', which was then thought to be a medium essential for propagation of light. If 'ether' existed, then his interferometer would show a fractional fringe shift of 0.36, which appears to be ignorable. However, he claimed that his interferometer was accurate enough to record such a small fringe shift if it had really occurred. Repeated experiments at different places and in the different seasons showed no fringe shift, which left no doubt about the absence of hypothetic 'ether'. Thus light can travel in vacuum also. This was one of the concepts that became the basis for theory of relativity, credited to the contemporary Physicist, Albert Einstein. Albert Michelson was awarded the prestigious Nobel prize in Physics in 1907.

(Entirely optional)



1.4 MICHELSON'S INTERFEROMETER

Circular fringes, but where is the circular symmetry?

We now discuss an interferometer having extreme precision and versatility, which was designed by Albert Michelson. Consider a set up in the Figure 1.10. It is observed that ray incident on the beam splitter divides itself into two rays of approximately same amplitude. These

rays reunite at the point of their origin itself, when reflected by the Mirrors M1 and M2. If the Michelson's interferometer is aligned perfectly, then it produces circular fringes. Michelson's interferometer has a few important applications. One such application is the measurement of wavelength of monochromatic light.

Note that, both the rays 1 and 2 travel a double distance, from the point of their origin to the point of interference. As noted earlier the point of origin and interference is same. Now if M1 is moved up by $\lambda/2$, then the path difference between the ray 1 and 2 will increase by λ . Thus the point which originally corresponded to n^{th} fringe will now correspond to $(n+1)^{\text{th}}$ fringe. Thus $(n+1)^{\text{th}}$ fringe will replace the n^{th} fringe. In the similar manner $(n+2)^{\text{th}}$ fringe will replace $(n+1)^{\text{th}}$ fringe. Thus a crosswire focused on any fringe will encounter an inward shift of 1 fringe. This leads to a conclusion that if M1 is moved through a distance of X, then the path difference between ray 1 and 2 will increase by $2X$. If this corresponds to the shift of N fringes, then we can write

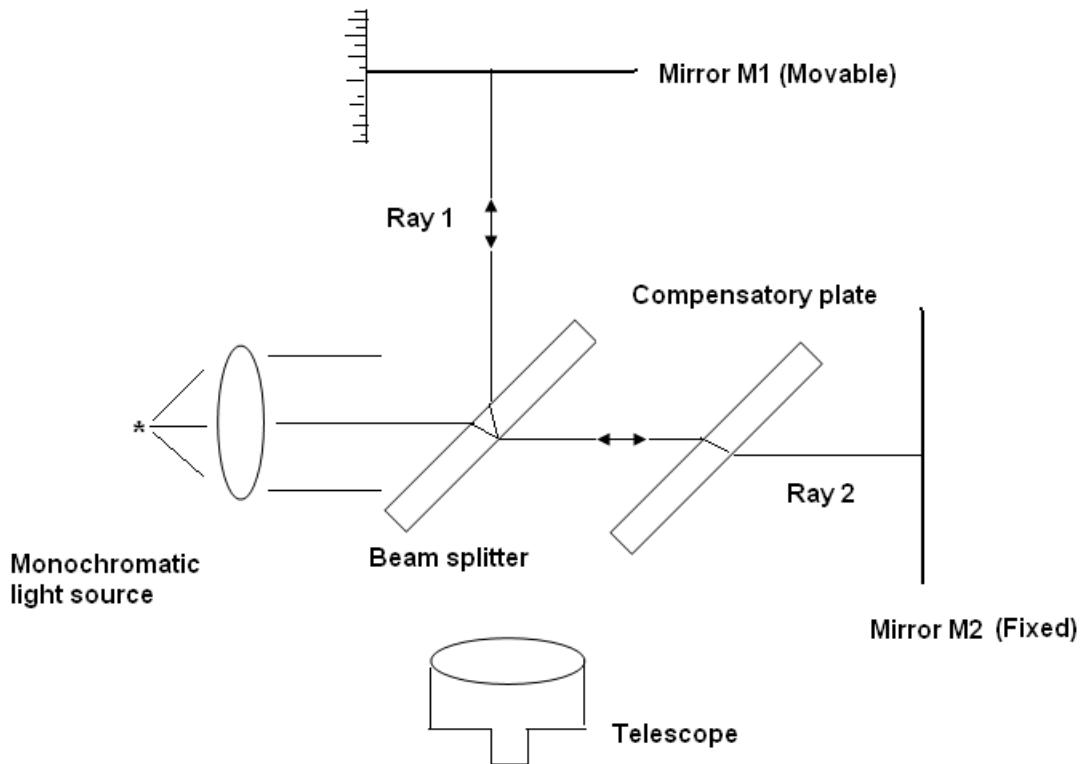


Figure 1.19 The schematic of the Michelson's interferometer

$$2X = N\lambda$$

... (1.26)

Above eqn, though appears simple, depicts two major applications of Michelson's interferometer. If the distance moved by M1, i.e. X is measured with the help of a screw gauge attached to it and if the shift in the fringes, N is counted then λ can be calculated. If λ is known then, on counting the shift N in the fringes, the unknown distance X moved by M1 can also be measured to a great precision. In the similar manner the change in X can also be measured by using $2\Delta X = N\lambda$. Indeed, the Michelson's interferometer plays a major role in the cases where the unknown distances or the change in the distance are too small to be measurable by conventional techniques.

If a transparent film of thickness t and refractive index μ is inserted in the path of any arm of the Michelson's interferometer, the owing to the denseness of the film, the path difference will increase by $(2\mu t - 2t)$ i.e. $2t(\mu - 1)$. This will also lead to shift of say N fringes. We know that one fringe shifts if the path difference increase by λ . Thus,

$$2t(\mu - 1) = N\lambda \quad \dots(1.27)$$

Above eqn emphasizes importance of the Michelson's interferometer. If λ is known and if the shift in the fringes after insertion of the film is counted, then the thickness t of the film can be calculated if its refractive index μ is known, otherwise if t is known then μ can be calculated. Eqn 1.27 is applicable to liquids, gases and air also. In such case t will represent the thickness of the cell in which these species are filled.

Michelson's interferometer can also be used to measure the resolution of the spectral lines from a bi-chromatic source. In this case, two sets of circular fringes will be observed, one corresponding to λ_1 and another to λ_2 . If, the bi-chromaticity of the source is extremely small, then λ_1 and λ_2 will have extremely closer values. Thus the fringes corresponding to these wavelengths will overlap. Of course, one to one overlap is never possible, as the wavelengths are not equal.. If the maxima in one fringe pattern overlap on the minima of another one, then the composite interference pattern will almost fade out. If the maxima of both the fringe patterns overlap, then the composite fringe pattern will show distinct fringes. Now consider a case where the pattern is distinct. Let the corresponding position of M1 be X_1 . Now if M1 is moved up, then the fringes in both the interference patterns will start getting shifted, but not at the same rate. According to eqn 1.24, the fringes corresponding to lower wavelength, say λ_1 will shift faster than λ_2 . Now if the movement of M1 is continued, then obviously the next stage will be corresponding to the disappearance of the fringes. If M1 still continues to move, then the fringe pattern will once again become distinct. Let the position of M1 corresponding to this distinct condition be X_2 . Considering the fact that M1 has moved such that the fringe pattern has shifted from a given distinct condition to the immediately the next one, we can write

$$2(X_2 - X_1) = 2X = n\lambda_2 \text{ and} \quad \dots(1.28)$$

$$2(X_2 - X_1) = 2X = (n+1)\lambda_1 \quad \dots(1.29)$$

Eqns 1.28 and 1.29 are simultaneous, thus

$$\begin{aligned} n\lambda_2 &= (n+1)\lambda_1 \\ n\lambda_2 &= n\lambda_1 + \lambda_1 \\ n(\lambda_2 - \lambda_1) &= \lambda_1 \\ n &= \frac{\lambda_1}{(\lambda_2 - \lambda_1)} \end{aligned}$$

Substituting n in eqn 1.29

$$2X = \frac{\lambda_1 \lambda_2}{(\lambda_2 - \lambda_1)} = \frac{\lambda_{av}^2}{(\lambda_2 - \lambda_1)} \quad \dots(1.30)$$

Thus, if λ_{av} is known by any other technique, then $(\lambda_2 - \lambda_1)$ can be calculated and the bi-chromatic source can be perfectly resolved. In fact, this technique also helps us to know, whether the source is monochromatic or bi-chromatic.

Albert Michelson carried out many enterprising experiments with his unique interferometer. One of these was to standardize a meter. For this purpose, he used a standard unit of length called as etalon. The shortest etalon had a length 0.390625 mm. M1 was moved from one end of the etalon to another one, and the shift in the fringes was counted. Then an etalon having its length double than that of the first etalon was experimented and the shift in the fringes was measured. In second case exactly double shift in the fringes was expected. Thus the average of experimentally measured fringes shifted in the case of first and the second etalon was calculated. Then, third etalon having its length double to that of the second etalon was experimented and the shift in the fringes was counted. An average of the shift in fringes of three successive etalons increased the accuracy by threefold. In the same way, etalons of increasing lengths were experimented. The length of last etalon was 10 times that of the first etalon. The average of the counted and the expected fringes increased the accuracy by tenfold. These measurements were then generalized to 1 meter. The results were

$$1 \text{ meter} = 1,553,163.5\lambda_R$$

(Where λ_R is the wavelength of the Cadmium red line = 6438.4722 Å⁰)

$$1 \text{ meter} = 2,083,372.1\lambda_B$$

(Where λ_B = the wavelength of the Cadmium blue line = 4799.9107 Å⁰)

And,

$$1 \text{ meter} = 1,966,279.7\lambda_G$$

(Where λ_G = the wavelength of the cadmium green line = $5085.8240 \text{ A}^\circ$)

Michelson also used his interferometer along with Edward Morely to verify the then hypothesized ether. The idea was that, as the arms of the interferometer had equal lengths and as they were perpendicular to each other, the distance traveled by the ray 1 and 2 in the ether passing in a direction opposite to the rotation of the earth would differ. This would result in a shift in the fringe pattern. Mathematics indicated that a fractional shift of 0.36 would occur. But this was not observed. The experiment was repeated at various places and in the various seasons, but in any case the expected fringe shift was not observed. Michelson claimed that his interferometer was accurate enough to record the fringe shift of 0.36 if it really occurred. The absence of any fringe shift convinced the Physicists about the absence of the ether. Thus light could travel in vacuums also. This conclusion convinced Albert Einstein to propose special theory of relativity. Michelson also measured the diameters of the stars to a great precision using his interferometer.

In Caltech, an advanced version of Michelson's interferometer is being installed. This interferometer proposes to detect the gravitational waves if they exist. The undulations created by the gravitational waves in the path of ray 1 and 2, would cause shift in the fringes. If this is observed then the existence of gravitational waves will receive an experimental confirmation. It may be noted that the length of each arm in this interferometer will be 4.0 km . The name of this lab is LIGO (Laser Interferometer Gravitational-Wave Observatory, <http://www.ligo.caltech.edu/>)

As mentioned in the beginning of this chapter, several interferometers exist, each having its own applications. One of them is Fabry-Perot Interferometer. However, their discussion is beyond the scope of the first year engineering course.

Example 1.21 Assume that the movable mirror in the Michelson's interferometer is moved through the L.C. of the micrometer screw gauge attached to it, which is $10 \mu\text{m}$, i.e. 0.001 cm . If the interferometer is operated using sodium light having wavelength 5890 A° , then how many fringes will be shifted?

Solution: We have

$$\begin{aligned}2X &= N\lambda \\2 \times 0.001 &= N \times 5890 \times 10^{-8} \\N &= 0.02945\end{aligned}$$

This is impractical answer. Thus, we should assume realistic data

Example 1.22 Through what distance the movable mirror in the Michelson's interferometer should be moved so that at least one fringe is shifted? The source used is sodium having wavelength 5890 \AA .

Solution:

$$2X = N\lambda$$

We have

$$2X = 1 \times 5890 \times 10^{-8} \text{ cm}$$

$$X = 29.45 \mu\text{m}$$

$$X = 0.02945 \text{ mm}$$

Considering that the L.C. of the micrometer screw gauge is 0.001 cm, we need to approximate the above answer to $X = 0.0295 \text{ mm}$.

Compulsory



1.4 INTERFERENCE COATINGS

How to minimize the unwanted reflection or transmission of light

A camera can produce better images if it receives maximum light from the object to be photographed. Solar cells can also give better electrical output if maximum light passes in. But this is prohibited by the reflection of light. For ex. in camera, the reflection of the light from the lens is nearly 4% and thus it reduces the transmitted intensity to 96%. Can this problem be solved using the interference of thin films? Further, as we may have experienced, certain glasses, especially those used in the windows or the doors of the commercial shops, or windows of automobiles or sometimes in buildings are excessively reflective, but only from one side. The excessive reflectivity doesn't allow the observer to look on the other side. While, as the reflectivity is only from one side, the observer at the non-reflective side can look through the glass. The rhinestones can also be provided with attractive looks by using such techniques. The reflectivity of the mirrors can also be enhanced to its ideal value (100 %). Interference filters which can pass an extremely small narrow band of the wavelengths can also be designed using such concepts. In short, what we would like to discuss here are Anti-Reflection Coatings (ARC) and the High Reflection Coatings/Anti-Transmission coatings (HRC/ATC).

Consider the Figure 1.10

We start with the usual eqn. of thin film interference,

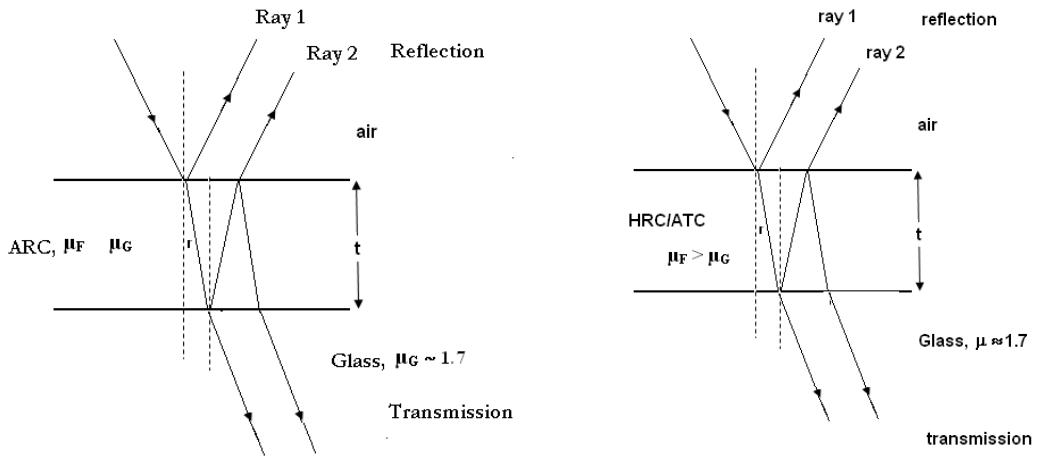


Figure 1.10 Interference coatings (a) ARC and (b) HRC/ATC

$$PD_{R,I,II} = 2\mu t \cos r \pm \frac{\lambda}{2} \quad \dots(1.31)$$

In general, the incident ray falls very close to the normal, thus $r \rightarrow 0$. Note that in the case depicted in Figure 1.9, the Stoke's factor is absent in the above eqn. A destructive interference between reflected rays suppresses the reflection. Consequently the transmission enhances. Thus

$$PD_{I,II} = 2\mu t \cos r = (2n \pm 1) \frac{\lambda}{2} \quad \dots(1.32)$$

If the film is thin enough, we can safely assume $n \approx 0$. Thus

$$\boxed{t_{ARC} = \frac{\lambda}{4\mu}} \quad \dots(1.33)$$

Thus, a film having thickness t , which satisfies above eqn behaves as ARC. ARCs are made of MgF_2 ($\mu = 1.38$), SiO_2 ($\mu = 1.49$).

The discussion of HRC/ATC and ARC is not exactly opposite. We just need to replace the film in ARC by a material denser than the substrate. The substrate is generally the glass. In such case, the Stokes's factor $\lambda/2$ is present in the eqn 1.31 (as suggested by Stoke's law). Further, a constructive interference between the reflected rays will make the film more reflective and less transmittive. Thus, we have

$$\begin{aligned} PD_{I,II} &= 2\mu t \cos r \pm \frac{\lambda}{2} = n\lambda \\ 2\mu t \cos r &= (2n \pm 1) \frac{\lambda}{2} \end{aligned}$$

For thin films, it is safe to assume $n \approx 0$. Generally, the incident ray passes close to normal. Thus $r \rightarrow 0$ and we get

$$t_{HRC/ATC} = \frac{\lambda}{4\mu} \quad \dots(1.34)$$

If t satisfies above eqn, the coating behaves as HRC/ATC. HRC/ATC does not work in opposite direction, as in such case when light passes through the glass, its wavelength, λ will decrease substantially ($\lambda' = \frac{\lambda}{\mu}$) and Eqn.(1.34) will not work for the decreased wavelength. Generally the materials used for HRC/ATC are TiO_2 ($\mu = 2.4$), ZnS ($\mu = 2.32$). By using multi-layer coatings, it is possible to block the transmission of all the wavelengths except one. Thus an extremely narrow band of wavelengths (with $\lambda \approx 11 \text{ \AA}$) can be transmitted. Such coatings are called as interference filters. Note that the eqn (1.33) for ARC and eqn (1.34) for HRC/HTC appear same, but they are not. We also observe that in both the cases t depends upon λ



Camera lens with and without antireflection coatings



Rhinestones coated with high reflection coatings. Look at the bright colors



Window of a car coated with high reflection coatings



Glasses coated with high reflection coatings are also used in buildings

RAPID REVIEW

Interference is not just a natural phenomenon responsible for colors in the oily films and the soap

bubbles. It forms the basis of instruments with extreme precision called as interferometers. Both the interferometers that we have gone through i.e. Newton's and Michelson's interferometer are based on thin film interference. Thin film itself is an interferometer. There are two kinds of thin films as mentioned below.

Thin film of constant thickness,

$$PD = 2\mu t \cos r \pm \frac{\lambda}{2}$$

Wedge shaped film

$$PD = 2\mu t \cos(r + \alpha) \pm \frac{\lambda}{2}$$

Thin films can produce two kinds of fringes; one is Fizeau's fringes, where t is varied and r is kept constant and another is Haidinger's fringes, where t is kept constant and r is varied.

Newton's interferometer involves a wedge film with circular symmetry. The measurement of diameter of Newton's rings leads to three applications given below.

$$R = \frac{\mu(D_m^2 - D_n^2)}{4((m-n))\lambda} \text{ and } \mu = \frac{(D_m^2 - D_n^2)}{(D_m^2 - D_n^2)}$$

Michelson's interferometer surpasses the Newton's interferometer in many respects. It is more versatile, more accurate and more sensitive. Its few applications are based on the following eqns

$$2X = N\lambda, \quad 2t(\mu - 1) = N\lambda \quad \text{and} \quad 2X = \frac{\lambda^2}{(\nabla\lambda)}$$

Interference also forms the basis of ARC and HRC/ATC, the formula is $t_{ARC/ATC} = \frac{\lambda}{4\mu}$. For

ARC, the coating is rarer than the substrate, while for HRC/ATC is the case is opposite.

APPENDIX I TO CHAPTER 1

The Optical Path

Consider Figure 1.11, which is self-explanatory

We have

$$v = \frac{d}{t}$$

$$v \times t = d$$

$$\frac{c}{\mu} \times t = d$$

$$ct = \mu d$$

Thus Optical path = $\mu \times$ Geometrical path

Thus optical path represents the distance travelled by the light with speed c , in a time t , which the light spends while travelling through a medium of thickness d with lowered speed v . In short, a path difference is created due to slow speed of the light in the denser medium. As an alternative explanation, we have

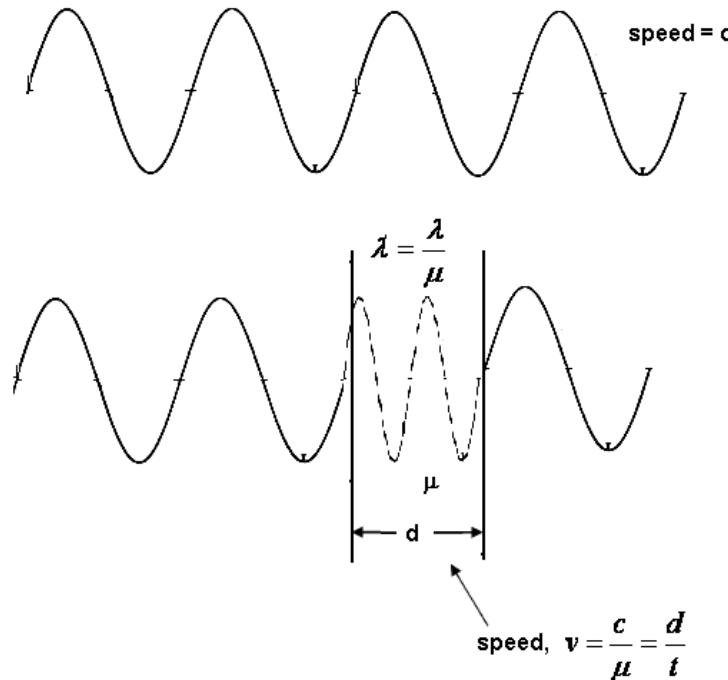


Figure 1.11The concept of Optical path

$$\begin{aligned}\lambda' &= \frac{\lambda}{\mu} \\ \lambda' \times \mu &= \lambda \\ n(\lambda' \times \mu) &= n\lambda \\ \mu \times (n\lambda') &= n\lambda \\ n\lambda &= \mu \times (n\lambda')\end{aligned}$$

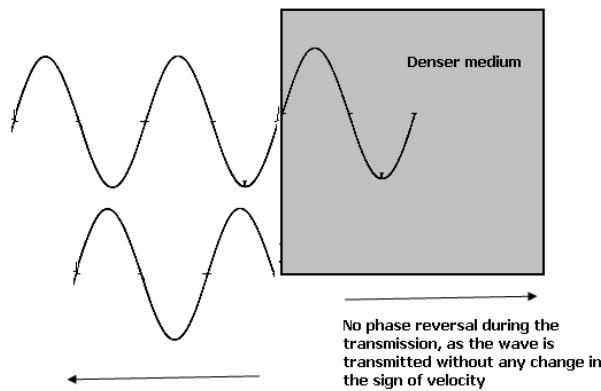
$$\text{Optical path} = \mu \times \text{Geometrical path}$$

Note that owing to the decreased speed in the denser medium, light carries a smaller wavelength λ' , while passing through the denser medium

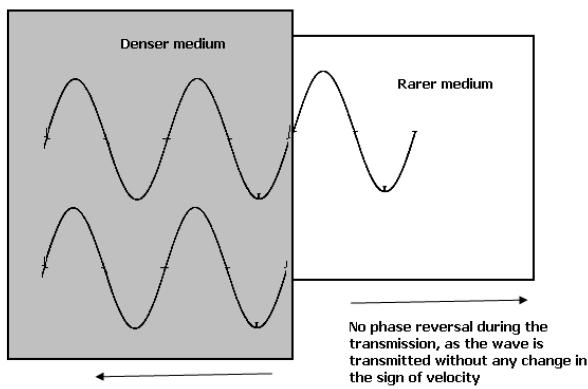
APPENDIX II TO CHAPTER1

The Stoke's Law

1. If a wave suffers a reflection from denser to rarer medium its phase is reversed
2. No phase reversal occurs during the transmission.
3. No phase reversal occurs if the wave is reflected from rarer to the denser medium



Phase reversal during the reflection,
as the wave is reflected back with the change
in the sign of the velocity



No phase reversal during the reflection,
as though the wave reflected back with the opposite velocity, the
particle velocity on the interface is not reversed as the
particles in the rarer medium are free to oscillate

Figure 1.10 Physics behind the Stoke's law

APPENDIX III TO CHAPTER 1

Derivation of Path Difference in case of Wedge-shaped film

Consider the following figure. We have

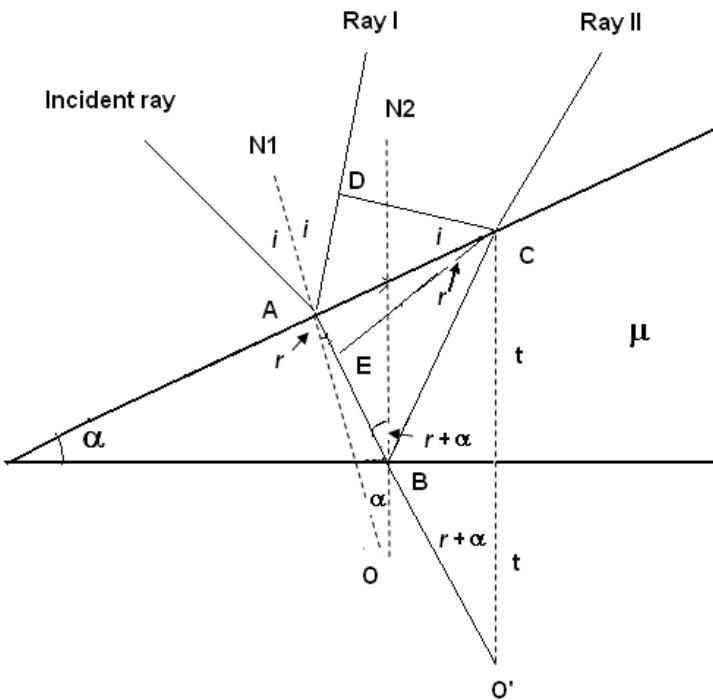


Figure 1.12 Ray diagram for the Wedge shaped film

$$PD_{I,II} = \mu(AE + EB + BC) - AD$$

$$PD_{I,II} = \mu(AE + EB + BO') - AD$$

$$PD_{I,II} = \mu(AE + EO') - AD$$

$$PD_{I,II} = \mu[AE + \cos(r + \alpha)] - AC \sin i$$

$$PD_{I,II} = \mu[AC \sin r + 2t \cos(r + \alpha)] - \mu AC \sin r$$

$$PD_{I,II} = 2\mu t \cos(r + \alpha)$$

With Stoke's law

$$PD_{I,II} = 2\mu t \cos(r + \alpha) \pm \frac{\lambda}{2}$$

QUESTIONS

It is only the first step that takes the effort. –Marquise du Deffand

General

Fundamentals of Optics

1. Consider any optical instrument, say the Human eye, and identify any three laws of Optics involved in its functioning.
2. Indicate how the design of such instrument requires precision optical engineering.
3. Mention any one application of the optics in the daily life.
4. Light is an electromagnetic wave. What is an electromagnetic wave?
5. According to particle character of the light, it consists of photons. What do you mean by a photon?
6. What do you mean by coherence?
7. Name as many monochromatic sources as known to you.
8. Name the best coherent source known to you.
9. Name as many polychromatic sources as known to you.

Basics of Interference

10. Interference explores the wave character of light. How?
11. Interference does not reveal whether light is longitudinal or a transverse wave. How?
12. What do you mean by a steady state interference pattern?
13. What conditions need to be satisfied for generating a steady state interference pattern?
14. What do you mean by a sharp interference pattern?
15. What conditions need to be satisfied for generating a sharp interference pattern?
16. The generation of a well-defined interference pattern requires a systematic and gradual variation of the path difference. What will happen if the path difference varies randomly?

1.1 Thin Film Interference

17. A thin film needs to be thin, why?
18. A thin film satisfies all conditions for generating a well-defined interference pattern except one. What is it? Why?
19. A film excessively thinner or thicker than the wavelength of the light cannot produce interference pattern. Why?
20. The fringes in case of a thin film of uniform thickness exposed to a point source are circular. Why?
21. The fringes in case of a thin film of gradually increasing thickness exposed to a broad source are straight. Why?
22. A thin film can produce the interference pattern, when either its thickness is varied by keeping the angle of incidence same, or the angle of incidence is varied by keeping thickness same. Is it possible to produce a well-defined interference pattern with the

combination of both these situations at once? Why? Why not?

23. Is it really necessary to expose the thin film to a monochromatic source in order to produce an interference pattern? Explain.
24. Evaluate whether the Stokesfactor will be present or absent in the following situations.
Also explain, why
 - a. A thin film existing in between denser media
 - b. A thin film existing in between rarer media
 - c. Medium above the film is denser and the medium below the film is rarer
 - d. Medium below the film is rarer and the medium above the film is denser
25. If a film appears dark from the reflected side, then at the same point it appears bright when viewed at the same angle but on the transmitted side. Is it ever possible to create a situation, where this does not happen? Why? Why not?
26. Can a thin film of randomly varying thickness and randomly varying refractive index produce an interference pattern? Why? Why not?.
27. A film excessively thinner or thicker than the wavelength of the light cannot produce interference pattern. Why?
28. The fringes in case of a thin film of uniform thickness exposed to a point source are circular. Why?
29. The fringes in case of a thin film of gradually increasing thickness exposed to a broad source are straight. Why?

1.2 Wedge shaped films

30. Why it is necessary for the wedge shaped film to have a very small wedge angle?
31. Can a wedge film produce a well-defined interference pattern when exposed to a white source? Why? Why not?
32. The formula for the fringe width of the wedge shaped film in this chapter is an approximate formula. Why?
33. Can a wedge film produce a well-defined interference pattern, if exposed to a broad source? Why? Why not?
34. A wedge shaped film can be conveniently used to inspect the diameters of the thin wires. How?

1.3 Newton's Rings

35. Newton's rings are concentric and circular. Why?
36. Newton's rings become thinner, as one moves away from the center. Why?
37. Newton's rings disappear if the Plano-convex lens is kept on the mirror instead of the glass plate. Why?
38. The Plano-convex lens used for producing the Newton's rings should have small curvature. Why?
39. How curvature and radius of curvature are related to each other? Are they directly proportional or indirectly proportional to each other?
40. Can Newton's rings be observed using bio-convex lens? Why? Why not?
41. Newton's rings show an outward shift if the lens is lifted up. Why?

42. Sometimes, Newton's rings show a bright center instead of dark and sometimes vice versa. Why?
43. If Newton's rings are observed with white source, then they appear colored. Why?
44. The colored Newton's rings exist only near the center. Why?
45. Newton's rings shrink in diameters when air is replaced by a denser medium. Why?
46. For better accuracy, the diameters of at least two rings are measured instead of one. Why?
47. The center of Newton's rings is not pinpointed, but is bulged. How?
48. The perfection in the devices like lenses and glass plates can be tested with Newton's interferometer. How?
49. Newton's rings cannot resolve a bi-chromatic light. Why?
50. Newton's rings disappear if the plano-convex lens is kept on the mirror instead of the glass plate. Why?

Optional



1.4 Michelson's Interferometer

51. In what ways Michelson's interferometer is better than the Newton's interferometer?
52. The lengths of the arms in the Michelson's interferometer should not be equal. Why?
53. Is it really necessary for the Michelson's interferometer to have compensatory plate? Why? Why not?
54. Does the accuracy of the Michelson's interferometer depend upon the least count of the screw gauge? Why? Why not?
55. What is the smallest length or the change in length measurable by Michelson's interferometer?
56. Michelson's interferometer is considered as a precision instrument for measuring dimensions or the change in dimensions. Why?
57. If the 'ether' existed, then the fractional fringe shift encountered in Michelson's Morley experiment would be 0.36. What do you mean by a fractional fringe shift? How can it be measured?
58. For the standardization of meter, Michelson used cadmium source. This source is polychromatic. How can a meter be standardized by using a polychromatic source?
59. What is the smallest resolution of a bi-chromatic source that can be resolved by using Michelson's interferometer?
60. Michelson's interferometer can resolve sodium source, but Newton's interferometer cannot. Why?
61. The sensitivity of the Michelson's interferometer increases with the length of the arms. How?
62. List any two interferometers other than Newton's interferometer and Michelson's interferometer.
63. Michelson's interferometer can be used to
 - e. Inspect the displacement in the walls of the Dam
 - f. Young's modulus of glass or metals

- g. Coefficient of the thermal expansion of glass or metals

Identify any one application of your choice and indicate how it is possible

Compulsory



1.5 Interference Coatings

64. Antireflection coatings (ARC) enhance the transmission of light just from 96 % to ~ 100%. Is there really any advantage?
65. The ARC is made up of materials denser than the substrate, while HRC/ATC is made of materials rarer than substrate. Can this be made possible in opposite manner? How?
66. Lenses of most of the cameras appear purplish. Why?
67. The purplish tint is not observed if viewed at the different angels. Why?
68. The rhinestones can be made more attractive with HRC/ATC. How?

PROBLEMS

The significant problems of our time cannot be solved by the same level of thinking that created them.

- Albert Einstein

1.1 Thin film interference

1. Calculate the range of the angles in which the thin film having a thickness $0.55 \mu\text{m}$ being viewed at 68° will appear greenish. The wavelength of the green color varies from $4950-5700 \text{ \AA}^\circ$.
2. At what angle should film in the problem 1 be viewed, so that it appears reddish? The wavelength of red light is 6320 \AA° .
3. If the film is suspended on a dry road, it does not appear colored. Why?
4. What color, the film in the problem 1 will show, if viewed at the same angle, but from the transmitted side?
5. The range of visible light is $3800-4500 \text{ \AA}^\circ$ (violet), $4500-4750 \text{ \AA}^\circ$ (blue), $4760-4950 \text{ \AA}^\circ$ (cyan), $4950-5700 \text{ \AA}^\circ$ (green), $5700-5900 \text{ \AA}^\circ$ (yellow), $5900-6200 \text{ \AA}^\circ$ (orange), $6200-7500 \text{ \AA}^\circ$ (red). Can this range be resolved by a thin film having thickness $0.55 \mu\text{m}$? Why? Why not? Assume that the film has refractive index of 1.35 and is denser than road on which it is spread.
6. Which two colors can just be resolved by thin film of thickness 1389 \AA° and refractive index 1.35 at 45° ? Assume that the thin film is denser than the substrate.
7. What should be the minimum thickness of the thin film which just resolves violet and the blue color at 45° . Given the refractive index of the film is 1.35 and it is spread on the substrate which is denser than the film.

1.2 Wedge shaped films

8. Calculate the smallest dimensions that can be measured by using the technique of wedge film. Assume that the source used is sodium, having wavelength 5890 A° and the object is kept at a distance of 10.0 cm from the wedge. The L.C. of the traveling microscope used to measure the fringe width is $100 \mu\text{m}$ (0.01 mm).

1.3 Newton's rings (Newton's interferometer)

9. If the L.C. of the traveling microscope used in the Newton's interferometer is $100 \mu\text{m}$, then what is the smallest radius of curvature of the Planoconvex lens that can be measured? The medium involved in the set up is air.
10. It is known that the refractive index of the sugar solution increases with its concentration. When the Newton's rings set up is kept in a sugar solution with 40 % concentration having refractive index 1.40, the diameter of a certain ring is measured to be 0.290 cm. Now, when the 40 % sugar solution is replaced by 68% sugar solution, the diameter of that ring acquires a value of 0.285 cm. What is the refractive index of the 68% sugar solution?
11. Newton's rings are being observed with the help of a planoconvex lens of radius of the curvature 100.0 cm and the sodium light having wavelength 5890 A° . The set up is kept in the air. Calculate the diameter of the 5th dark ring in such case. We know that calculators can provide the answers up to several decimal places. However, if the L.C. of the traveling microscope used in the interferometer is $100 \mu\text{m}$, up to what decimal place the answer needs to be rounded off?
12. The diameter of the 5th dark ring measured using Newton's interferometer in air is 0.292 cm. Considering that the L.C. of the traveling microscope is not more than $100 \mu\text{m}$, What is the smallest refractive index measurable by using Newton's interferometer?
13. Assume that the Planoconvex lens and the glass plate are kept in a system which can be evacuated. Initially the system contains air, the refractive index of which is accurately known to be $\mu = 1.00027717$. The diameter of the 5th Newton's ring is measured to be 0.592 cm. Assume that the interferometer is operated using sodium light having wavelength 5890 A° . Now the air is slowly evacuated. To what extent the diameter of 5th Newton's ring will change? Will it increase or decrease?
14. The diameter of the 5th dark ring in presence of air is 0.592 cm. When the air is slowly evacuated, the diameter decreases to 0.591 cm. Calculate the refractive index of the air.

Optional



1.4 Michelson's Interferometer

15. Assume that you are operating the Michelson's interferometer with the cadmium source emitting red, blue and green light having wavelengths $6438.4722 \text{ A}^\circ$, $4799.9107 \text{ A}^\circ$ and $5085.8240 \text{ A}^\circ$ respectively. If the movable mirror in each case is moved through the same distance then will the number of fringes shifted change in each case?

16. In above problem the wavelengths of cadmium are specified to the fifth decimal place. For ex. the wavelength of the blue line in cadmium is $4799.9107 \text{ \AA}^{\circ}$. Can such accuracy be achieved using Michelson's interferometer? Verify
17. How many fringes will be shifted if the movable mirror in the Michelson's interferometer is moved through a distance of 0.0295 mm. The source used is sodium-having wavelength 5890 \AA° . Sodium is a bi-chromatic source, and its next wavelength is 5896 \AA° . How many fringes will be shifted in case of second wavelength? Comment on the result
18. Do you feel that the distance moved by the mirror and hence, the fringes shifted are too many to distinguish between the sodium line? Try the smallest possible distance so that only one fringe is shifted.
19. What is the smallest length or the change in the length that can be measured using, Michelson's interferometer, if operated by He-Ne laser having wavelength 6328 \AA° .
20. Calculate the smallest thickness of the thin film that can be measured by using Michelson's interferometer. The refractive index of the thin film is 1.38 and the source is He-Ne laser having wavelength 6328 \AA° .
21. A glass cell having width 10.0 mm filled with air is inserted in the path of one of the interfering beams. The interferometer is operated with laser having wavelength 6328 \AA° . As the air is slowly evacuated the fringes start getting shifted. In which direction the fringes will shift? After the complete removal of the air 8.76 fringes are shifted. What is the refractive index of the air?
22. Can the thickness of the glass cell in the above problem be selected arbitrarily? Why? Why not?
23. Consider a cell having thickness 1.00 cm filled with water having refractive index 1.33. The interferometer is operated using laser light 6328 \AA° . The water is removed slowly. How many fringes will be shifted when water is removed completely from the cell?
24. Michelson's interferometer is used to resolve the sodium doublet, with an average wavelength of 5893 \AA° . The movable mirror is moved through a distance of 0.2894 cm, so that, the fringe pattern moves from one distinct condition to the immediately the next one. What is the resolution of the spectral lines? Also identify the wavelengths of both the spectral lines.
25. What is the smallest resolution of the spectral lines that can be measured by the Michelson's interferometer? Assume that the average wavelength is 5983 \AA° .
26. For standardization of the meter Michelson used Cadmium source and redefined the meter in terms of various wavelengths of cadmium. This requires separation of the spectral lines of the cadmium source. How to separate the colors of cadmium line?
27. While redefining the meter in terms of the spectral lines of Cadmium, Michelson uses the values of the spectral lines up to fifth decimal place. Is it ever possible to measure the spectral lines of any source up to fifth decimal place? Why? Why not?
28. The movable mirror of M.I is attached with the wall of the Dam. In 365 days, 3650 fringes are shifted. What is the net displacement in the wall of the Dam?
29. If the wall of the Dam is attached to any mirror, say movable or fixed and if the entire M.I. is perfectly clamped and aligned then it is possible to siren when the earth is about to quake. How?
30. The M.I. in LIGO (Laser Interferometer Gravitational-Wave Observatory has its arms 4 km long. Why such longer arms?

31. In LIGO, the beam has to cover a back and forth distances of $4\text{ km} + 4\text{ km} = 8\text{ km}$. can any beam retain its intensity after travelling such a large distance?
32. Refractive indices of a few gases are provided below. How many fringes will be shifted, if the cell containing vacuum at first is completely filled with the gases? M.I. is being operated with Laser having wavelength 6328 \AA . Do you feel that the measurements will be affected by the density of the gases being filled? And what about the impurities in the gases? The suggested thickness of the measuring cell is 1.00 cm

- h. Ammonia ($\mu = 1.000\ 376$)
- i. Carbon dioxide ($\mu = 1.000\ 449$)
- j. Carbon monoxide ($\mu = 1.000\ 338$)
- k. Chloroform ($\mu = 1.001450$)

Compulsory



1.5 Interference coatings

33. An ARC made up of MgF_2 having refractive index 1.38 is to be coated on the lens of a camera. The glass used in the lens is crown glass having refractive index 1.52. the incidence is normal. The brightest color in the visible spectrum is green and its wavelength is 5500 \AA . What should be thickness of the coating?
34. A rhinestone is to be made attractive by coating its seven sides reflective with seven different ATCs so that those seven sides show seven different colors brightly. The material used for the coating is TiO_2 having refractive index 2.4. The wavelengths of the different colors in the visible spectrum are given below. Assume normal incidence

Color	Wavelength
violet	380–450 nm
blue	450–475 nm
cyan	476–495 nm

green	495–570 nm
yellow	570–590 nm
orange	590–620 nm
red	620–750 nm

REFERENCE BOOKS

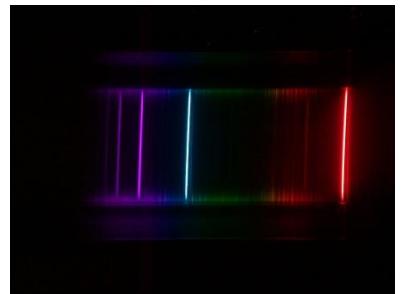
1. Fundamentals of Physics Extended, David Halliday, Robert Resnick, Jearl Walker,, John Wiley & Sons
2. The Feynman Lectures on Physics (3 Volume Set), by Richard Phillips Feynman (Author), Robert B. Leighton (Contributor), Matthew Sands (Contributor), The New Millennium Edition, Pearson Education India
 - Excellent websites on this book
 - i. www.feynmanlectures.caltech.edu/
 - ii. www.feynmanlectures.info/
3. A Textbook of Engineering Physics, M N Avadhanulu & P G Kshirsagar, 10th Edition, S. Chand and Company
4. Fundamentals of Optics, by Francis Jenkins, Harvey White , Tata Mcgraw Hill Publishing Co Ltd
5. Optics, Ajoy K. Ghatak, 5th Edition, McGraw Hill Education,
6. Optics, Eugene Hecht, 4th edition, Addison-Wesley
7. M. Born and E. Wolf, Principles of Optics, Cambridge University Press
8. A Text Book Of Optics, Brijlal, Dr. N. Subrahmanyam, Dr. M. N. Avadhanalu, 25th Edition, S. Chand and Company
 - i.

WORLD WIDE WEB

1. <https://www.photonics.com/>
2. SPIE - the international society for optics and photonics: spie.org/
3. Optical Society of America (OSA): <http://www.osa.org/>
4. Optical Society of India: www.osiindia.org/

CHAPTER 2

Diffraction



The first two images on the left indicate an advantage of the Diffraction Grating over prism. The Diffraction grating disperses the light into its spectrum, but with higher resolution and higher dispersion than prism. On the right side, a few of the antennas in Giant Meter-wave Radio Telescope (GMRT) can be observed. The GMRT, which is situated near Pune, consists of 30 parabolic dishes, each having 45 m diameter. The angular resolution of GMRT is nearly 2 to 60''. The resolution of any optical device including GMRT correlates with the physics of Diffraction Grating. How?

The answer to this question is in this chapter

Index

2.1 INTRODUCTION

Diffraction is not just the bending of light

2.2 SINGLE SLIT

Basis of multiple slits

2.3 DOUBLE SLIT

The diffraction and interference effects are unified

2.4 MULTIPLE SLITS:

How to produce bright and sharp maxima

2.5 DIFFRCTION GRATING

A super prism, the basis of spectroscopy

2.6 RAYLEIGH'S CRITERION OF RESOLUTION

For bare resolution of the images, the central maxima of one image should overlap on the first minima of adjacent image and vice versa

2.7 DISPERSIVE POWER AND RESOLVING POWER OF A GRATING

A grating having good dispersive power may not have good resolving power and ... vice versa

2.8 RESOLVING POWER OF THE GRATING

Why gratings are made up of enormously large number of slits

2.9 HUMAN EYE, TELESCOPES, MICROSCOPES, CAMERA and BIONACULARS

Resolving power is governed by diffraction effects

2.1 INTRODUCION

Diffraction is not just the bending of light

In day-to-day life, we find diffraction taking place at several places and situations, such as shadows and the images with diffused border, our ability to hear a person across the door but not to see him, our ability to resolve the images, when they are close to us, the binoculars, cameras, movie projectors, telescopes, VLAs, multi-antenna RADARS, our ability to stretch the 3.00 mm lens in the eye a little bit, if we start losing the resolution. And, probably, GOD has gifted us, two eyes, two ears, one nose, but with two nostrils, only one tongue, but with millions of sensors on its surface, just because, He wants us to have best possible resolving power. (Why one mouth then? possible HE wants us to see more and hear more)

The most promising application of diffraction is diffraction grating –asuperprsim, which separates the colors of light with incredible dispersion and resolution. Further, diffraction correlates with the resolving power of all optical instruments from human eye to the telescope and binocular to the movie cameras.

A great mistake is made, when diffraction is defined just as bending of light. Figure 2.1 shows a typical intensity pattern consisting of maxima and minima when a slit having its width close to wavelength of light is illuminated by light. The wave-front flares out, and due to the interference of the waves emanating from the Huygens's secondary wavelets, re-originating from within the slits, a pattern of maxima and minima is created.

	<p>Augustine Jean Fresnel 1788-1827: Educated in Caen (France) and at Polytechnique He established (with Arago) that light is a transverse wave whose two orthogonal polarizations do <i>not</i> interfere with each other. He also presented a rigorous analysis of diffraction using wave theory of light. He also proposed the use of Fresnel lenses in the lighthouses. His explanation for the single slit diffraction using Huygens's theory of secondary wavelets is considered authentic.</p>
---	--

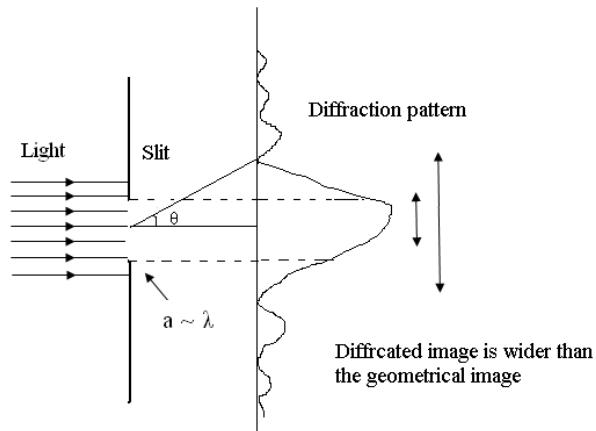
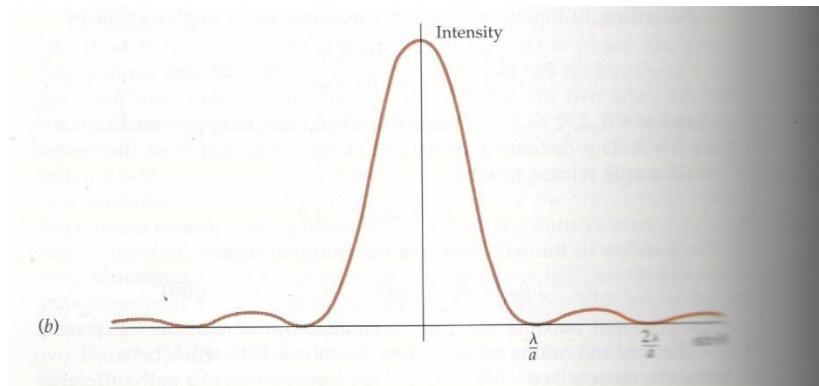
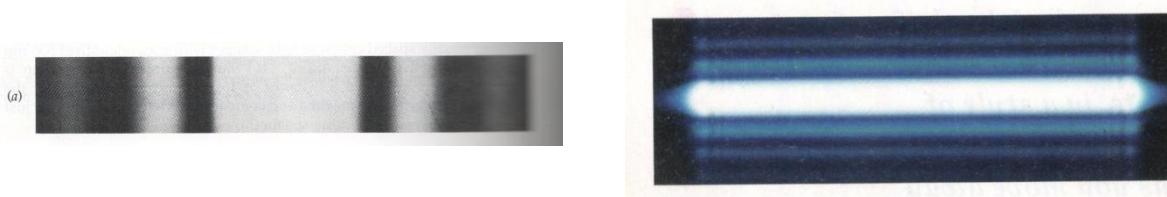


Figure 2.1 A typical description of diffraction. Diffraction is not just the bending of light

This is called as diffraction pattern. Thus diffraction involves obstruction of a wavefront by an obstacle, say a slit, production of Huygens secondary wavelets within the slit, production of secondary light waves from secondary wavelets from within the slits, interference of these secondary wavelets diffracted in various direction resulting in to a diffraction pattern.

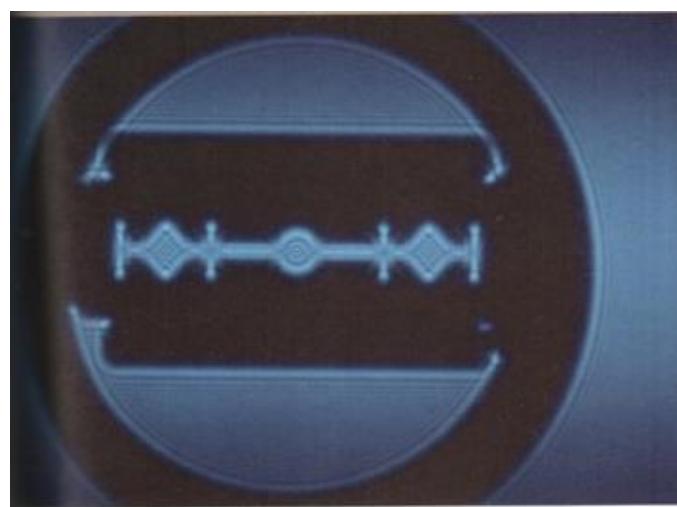


Single slit diffraction: Graphical

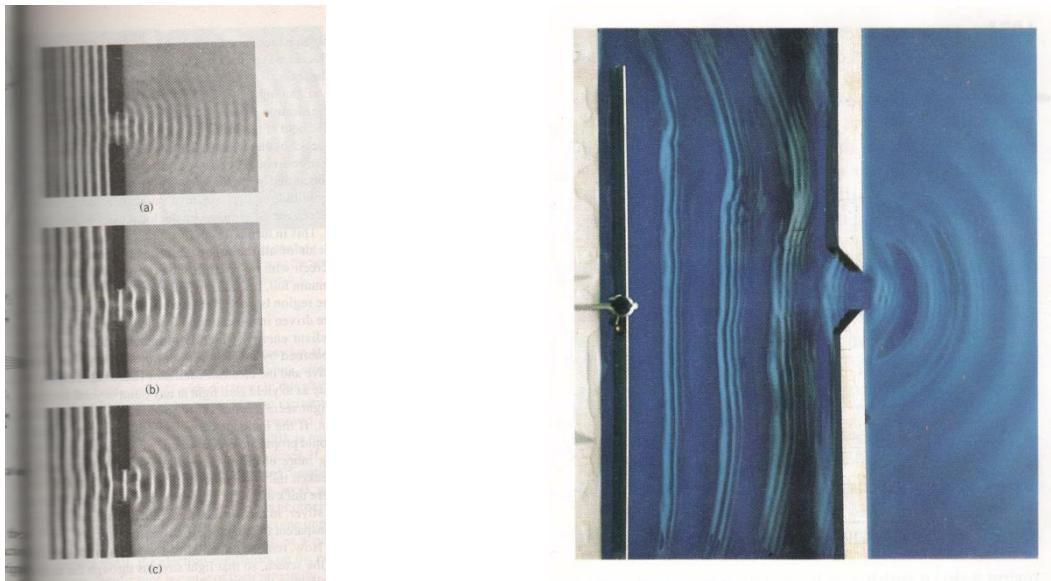


Single slit diffraction: Actual

Diffraction has two physical effects; one is widening of principle images and second is production of secondary images



Diffraction pattern of a blade: the images are not sharp due to diffraction



Diffraction of waves on the water surface. The wavefront flares out due to diffraction

Interference and diffraction

It is extremely difficult to identify the fundamental difference between interference and diffraction, as in both the phenomena, the superposition of secondary wavelets is involved. In an

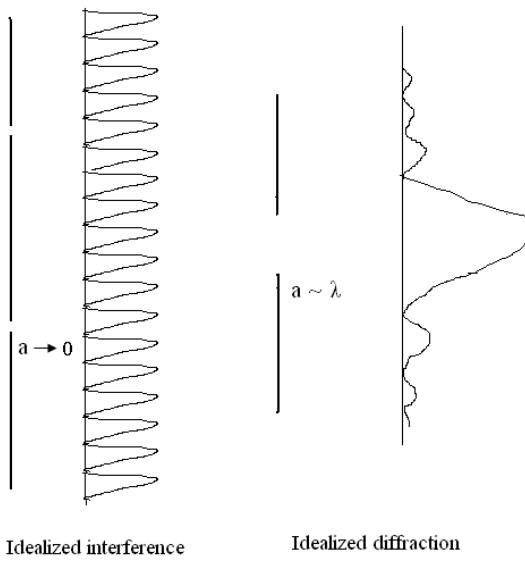


Figure 2.2 Idealized interference and diffraction. In reality, the cases can never be isolated

idealized interference, say in Young's double slit experiment; only two secondary wave-fronts are involved, which emanate from two point-like slits. In diffraction, the obstacle, say a slit, has

finite width and several secondary wavelets are formed within it. In idealized interference, as shown in Figure 2.2, the pattern consists of alternate maxima and minima of nearly same width and intensity, and in idealized diffraction, the pattern consists of principle maxima, surrounded by minima and secondary maxima. While it is practically impossible to have two point-like slits in interference, it is equally difficult to define the exact number of secondary wavelets in the slit involved in diffraction. Thus interference and diffraction represent the extreme cases of principle of superposition, the former involving limited number of secondary sources and the later involving extremely large number of secondary sources.

 Joseph von Fraunhofer (1787-1826)	Joseph Von Fraunhofer 1787 -1826 Once an undereducated apprentice, he established his own optical industry, where he designed and fabricated several devices and instruments such as loupes, prisms, microscopes, telescopes, astronomical reflectors etc. He worked with several contemporary Opticians. Fraunhofer described his investigations of light by gratings which were initially made by winding wires around parallel screws, a device called as diffraction grating. Indeed he was the first to make a diffraction grating. Using his grating he rediscovered almost 574 dark lines in the solar spectrum, which are called Fraunhofer lines.
---	---

Fresnel and Fraunhofer diffraction

In nature, the size of obstacles, and the nature of wave-fronts, as well as the nature of sources is beyond control. The diffraction involved in such cases may be treated as ‘natural’ diffraction and is referred as Fresnel’s diffraction. The wave-front falling on the obstacle can be made more systematic, say parallel, by using a lens and placing the source at its focal distance. This becomes a simplified case of the diffraction, because when the plane wave-front passes through the slit, the Huygens’s secondary wavelets generated in the slit act coherently in the beginning. This simplifies the analysis to a significant extent.

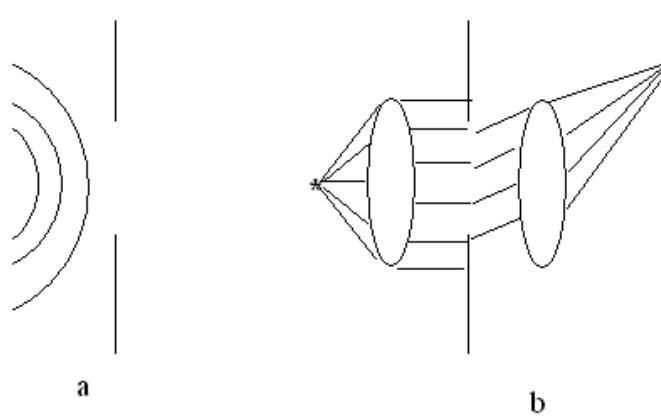


Figure 2.3 a. Fresnel's and Fraunhofer's diffraction

This case may be thus treated as ‘artificial/laboratory’ diffraction and is called as Fraunhoferdiffraction. Almost all the instruments related with diffraction are based on the Fraunhofer’s case.

2.2 SINGLE SLIT(**Compulsory, but the derivation is optional**)

A basis of multiple slits

The intensity at any point, say P in the diffraction pattern of a slit can be calculated by using Fresnel’s explanation of the diffraction. According to him, when a plane wavefront passes through the slit, every point in the slit becomes the source of Huygens secondary wavelets. The waves originating from these secondary sources are diffracted in all relevant directions, and they interfere due to a lens kept in between the slit and the screen. The interference of these waves results in to a pattern consisting of maxima and minima, which is called as diffraction pattern. By applying the principle of superposition, it is possible to formulate an equation describing the diffraction pattern. Though the number of waves superimposing at all the points are same, their resultant varies from point to point. This is so because the waves, although same in number, interfere with different path differences at different points. A simpler method for adding these waves involves laws of vector addition. Let each wave be represented by a vector called phasor. (*For the detailed understanding of concept of phasor, and phasor diagrams refer appendix I to chapter 2*). Thus N numbers of waves correspond to N number of phasors. As the amplitudes of the waves are same, the length of phasors is same. When the waves superimpose on the screen, the phasors join each other resulting in to phasor diagram. In such phasor diagrams the length of each phasor is proportional to the amplitude of the corresponding wave and the angle between the neighboring phasors is proportional to the phase difference between the corresponding waves. Though the number of phasors in all the diagrams is same, the shape is different at different points, as the waves superimpose with different path differences at the different points. Consider a typical case, say center of the diffraction pattern, where all the waves meet with zero path difference. The corresponding phasor diagram is straight and is shown in the Figure 2.4 (a). The result of this phasor diagram can be obtained by joining the bottom of the first phasor with the tip of the last phasor. As the phasor diagram in this case is a straight line, the length of the cord is same as arc. The resultant amplitude i.e. the resultant intensity is thus maximum. As one move away from the center, the phasor diagram becomes curved, and cord becomes smaller than the arc. Thus the resultant amplitude i.e. resultant intensity decreases. The curvature gradually increases with the increase in the angle of the diffraction. During this process, a stage comes where the curvature acquires the shape of a complete circle and the tip of the last phasor touches the bottom of the first phasor. The length of the cord, i.e. the resultant amplitude reduces to zero. This state corresponds to minima. The second minima in the diffraction pattern obviously occurs when the phasor diagram completes the circle next

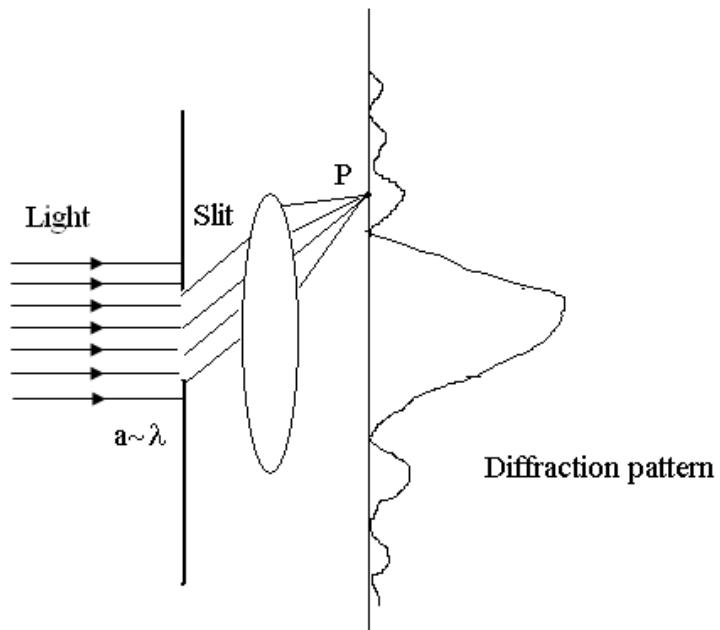


Figure 2.3 Single slit diffraction

Optional
↓

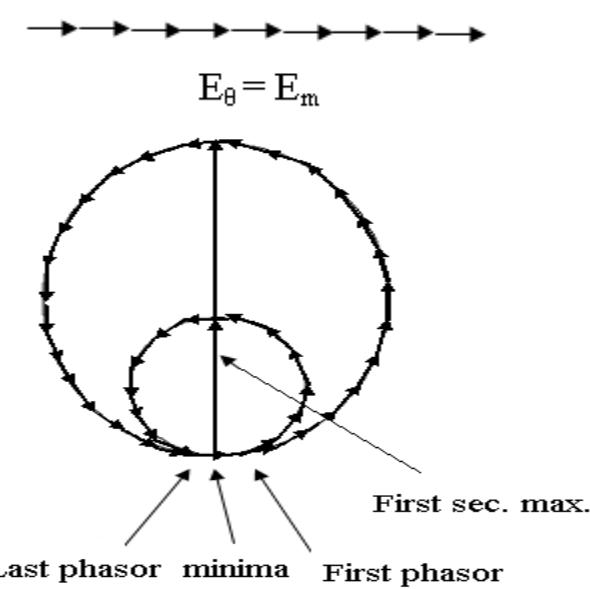


Figure 2.4 The various maxima and minima in the diffraction pattern can be explained with the help of phasor diagrams

Optional



time, however, now the circle has smaller diameter. This is so because the phasor diagram corresponding to the peak of the first secondary maximum is semicircular and this time the length of the cord i.e. the resultant vector needs to be smaller than that in the first case. Thus it is apparent that phasor diagram corresponding to any point can be drawn if and only if the phase differences between the phasors are known. The angles between the consecutive phasors correspond to the phase differences between the corresponding waves. These phase differences are the function of the corresponding angle of diffraction.

Consider a case, say point P, for which the angle of diffraction is θ . Thus all the secondary waves from the Huygens's secondary wavelets originating from the points within the slit, diffracted at θ will superimpose at point P. The construction of the phasor diagram at this point requires the phase differences between the consecutive waves, say. Though each point in the slit acts as the source of Huygens secondary wavelet, and though the distance between the neighboring points is zero, for the calculation of $\Delta\phi$ we need to assume that the main slit having width a is divided in to sub-slits, each having width, say ΔX . Each sub-slit is a point like sub-slit and thus though its width is ideally zero, it is only for exaggeration that we assume that ΔX has finite but extremely small value. Thus the center to center distance between the neighboring sub-slits is also approximately ΔX . From the figure 2.5, it appears that the path difference between zeroth and the first wave diffracted at θ is

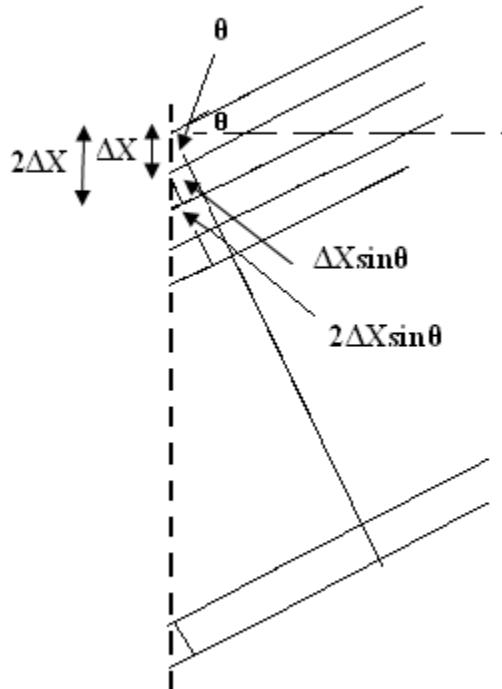


Figure 2.5 path difference between the waves diffracted at θ

$$PD_{01} = \Delta X \sin \theta$$

The corresponding phase difference between the zeroth and the first wave is

$$\nabla\phi = \frac{2\pi}{\lambda} \Delta X \sin\theta \quad \dots(2.1)$$

The path difference between the zeroth and the adjacent wave diffracted at θ is

$$PD_{12} = 2\Delta X \sin\theta$$

Thus corresponding phase difference will be $2\Delta\Phi$. The path difference between the zeroth and the last i.e. n^{th} wave will be

$$PD_{0N} = N\Delta X \sin\theta = a \sin\theta$$

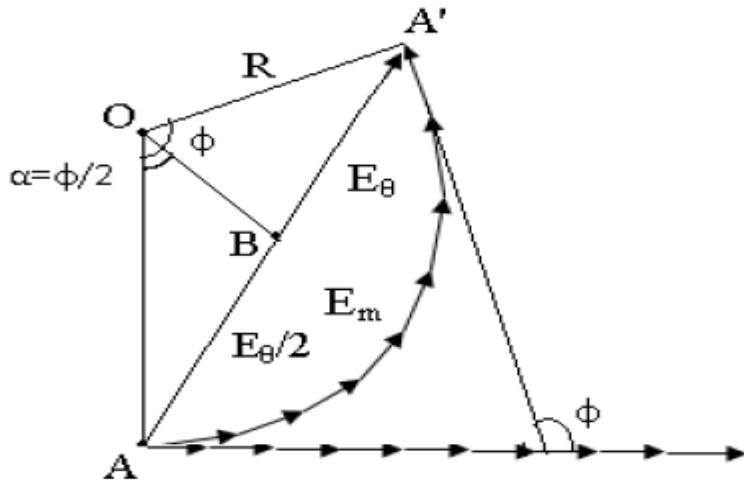


Figure 2.6 The phasor diagram at point P

Thus the phase difference between the zeroth and the last wave diffracted at θ is

$$\Phi = N\Delta\Phi = \frac{2\pi}{\lambda} \Delta X \sin\theta$$

Once $\Delta\Phi$ is known the phasor diagram can be constructed and is shown in figure 2.6. Note that in this diagram, the angle between the zeroth and the first phasor is $N\Delta\Phi$, the angle between the zeroth and second phasor is $2\Delta\Phi$. If the last i.e. n^{th} phasor is extended back and the first zeroth phasor is extended in forward direction then the angle between them is $N\Delta\Phi$ i.e. Φ . As the number of phasors in the diagram is extremely large, the diagram represents an arc rather than a polygon. Let the length of this arc be E_m (the suffix 'm' indicates 'maximum' length). A vector joining the bottom of the zeroth phasor to the tip of last phasor represents the resultant vector and its length represents the amplitude of the resultant wave, say E_θ (the suffix 'θ' indicates that the

resultant depends upon θ). The length of the normal drawn on the tangent to the arc at the end points (edges) represents the radius of the arc, say R . The angle between AO and A'O is Φ . Δ OBA represents the right angled triangle, and from its geometry

$$\frac{E_\theta}{2} = R \sin \frac{\phi}{2}$$

We know, $E_m = R\phi$, Substituting,

$$\frac{E_\theta}{2} = \frac{E_m}{\phi} \sin \frac{\phi}{2}$$

$$E_\theta = E_m \frac{\sin \frac{\phi}{2}}{\frac{\phi}{2}}$$

$$E_\theta = E_m \frac{\sin \alpha}{\alpha}, \quad \text{where } \alpha = \frac{\phi}{2} = \pi \frac{a}{\lambda} \sin \theta$$

Squaring both the sides

$$(E_\theta)^2 = (E_m)^2 \left(\frac{\sin \alpha}{\alpha} \right)^2$$

$$I_\theta = I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \quad \dots(2.2)$$

Equation 2.2 gives a general formula for calculating the intensity at any point having angle of diffraction θ . It appears that I_θ is the function of I_m , a/λ and θ .

The diffraction pattern has three typical intensities; central maxima, minima and secondary maxima. All these can be predicted by using equation 2.2

Eqns. (2.3) & (2.4) are compulsory, but their derivations are optional

Central maxima,

$$I_\theta = I_m \left[\frac{1}{\alpha} (\sin \alpha) \right]^2$$

$$I_\theta = I_m \left[\frac{1}{\alpha} \left(\alpha - \frac{\alpha^3}{3!} + \frac{\alpha^5}{5!} - \dots \right) \right]^2$$

Putting $\alpha = 0$ ($\Rightarrow \theta = 0$), we have $I_\theta = I_m$. Thus the center of diffraction pattern has maximum intensity

Minima

This requires $I_\theta = 0 \Rightarrow \sin\alpha = 0 \Rightarrow \alpha = m\pi$, where $m = \pm 1, \pm 2, \pm 3, \dots$ except 0

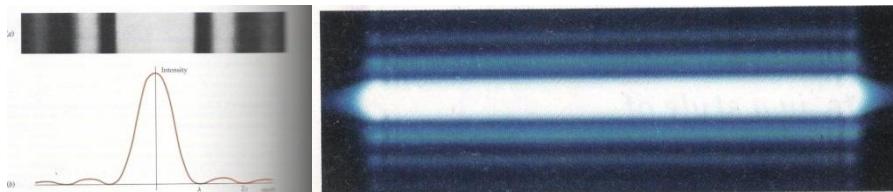
$$\begin{aligned} \alpha &= m\pi \\ \Rightarrow \pi \frac{a}{\lambda} \sin\theta &= m\pi \\ \Rightarrow a \sin\theta &= m\lambda \end{aligned} \quad \dots(2.3)$$

Equation 2.3 gives condition for minima. m^{th} minima occurs at those θ , which satisfy Eqn. (2.3)

Secondary maxima

Secondary maxima occurs approximately midway between consecutive minima. Thus when m is replaced by $(m + \frac{1}{2})$, we get condition for secondary maxima. Thus for secondary maxima

$$\begin{aligned} a \sin\theta &= \left(m + \frac{1}{2}\right)\lambda \\ \alpha &= \left(m + \frac{1}{2}\right)\pi \end{aligned} \quad \dots 2.4$$



Diffraction pattern of a single slit, graphical and Actual

All these problems are compulsory



Example 2.1 A slit having width $1.6933 \mu\text{m}$ is illuminated by sodium light having average wavelength 5893 Å° . Calculate the angular width of the central maximum. Also calculate the linear width of the central maximum if the screen is placed at a distance of 1 ft from the slit

Solution: The figure shows that the width of central maximum is 2θ , where θ is the position of the first minimum.

We have

$$\begin{aligned} a \sin \theta &= m\lambda \\ \theta &= \sin^{-1}\left(\frac{m\lambda}{a}\right) \\ \theta &= \sin^{-1}\left(\frac{1 \times 5893 \times 10^{-10}}{a 1.6933 \times 10^{-6}}\right) \\ \theta &= \sin^{-1}\left(\frac{1 \times 5893 \times 10^{-10}}{a 1.6933 \times 10^{-6}}\right) \\ \theta &= 20.37^\circ \end{aligned}$$

Angular width of the central maximum = $2\theta = 40.74^\circ$

From the figure $2L$ is the linear width of the central maximum and

$$L = 1 \text{ ft} \times \tan \theta$$

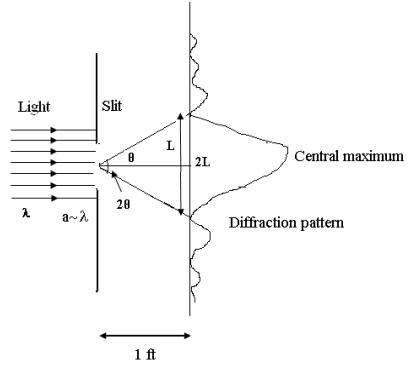
$$L = 1 \times 12 \times 2.54 \times \tan 20.37$$

$$L = 11.32 \text{ cm}$$

Linear width of the central maximum is = $2L = 2.63 \text{ cm}$

It is interesting to note that the linear width of the central maximum is significantly larger than the width of the slit. This property is useful in measuring the diameters of the thin wires, or monitoring them as they are being manufactured. Diffraction results in to widening of the images or the shadows. In case of extremely thin wires, it is easier to monitor the magnified shadow of the wire than the thin wire itself. Interestingly, thinner the wire, more it is difficult to monitor it directly, however, as the wire becomes thin, θ becomes large and the image or shadow becomes wider. We have already learnt, how wedge shaped films can be used to monitor the thin wires. Here, once again we note that the optical measurements are more convenient and precise than the mechanical measurements.

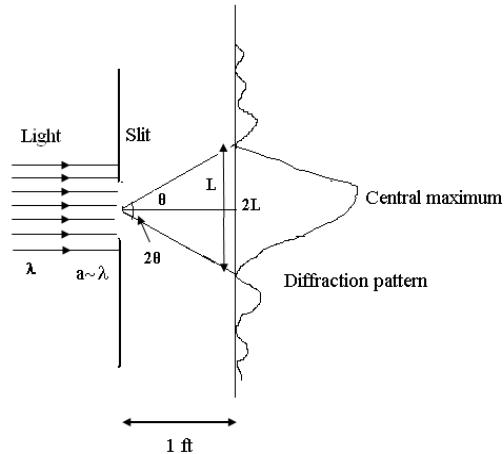
Example 2.2 Calculate the angular width of the central maximum, when slits having widths $a = \lambda$, $a = 2\lambda$ and $a = 20\lambda$ are used



Solution: Minima is given by

$$a \sin \theta = m\lambda$$

As the central maximum is bounded by first minimum on upper as well as lower side, we take $m=1$. Then for $a = \lambda$, 2λ , 20λ , we get $\sin \theta = 1, 0.5$ and 0.05 . The corresponding θ are $90^\circ, 30^\circ, 2.86^\circ$. The angular widths of central maximum keep on decreasing with the increase in the slit width. Note that its not absolute slits width, but the it is the ratio a/λ , which matters. We also understand that as $\frac{a}{\lambda} \rightarrow \infty, \theta \rightarrow 0$ indicative of the fact that diffraction effects become weak as the obstacle become larger than the wavelength of the wave being diffracted. This is the reason, why we can listen but cannot see the person standing behind the door. This is also the reason why bigger telescopes are better and why electron microscope is quite better than optical microscope. Further, this is also the reason, why X rays cannot be diffracted by optical diffraction gratings



Example 2.3 Calculate the angular positions of the first secondary maximum, when a slit of $1.6933 \mu\text{m}$ is illuminated by sodium source emitting two wavelengths 5890 A° and 5896 A° .

Solution

The position of the secondary maximum is given by

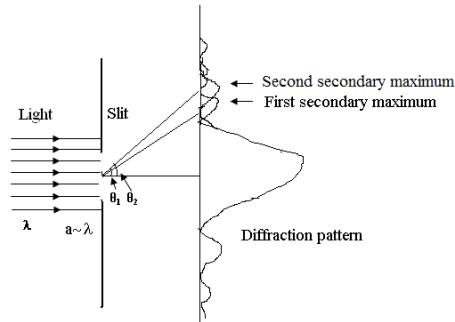
$$a \sin \theta = \left(m + \frac{1}{2}\right) \lambda$$

For first secondary maximum, $m=1$

For 5890 A°

$$1.6933 \times 10^{-6} a \sin \theta_1 = \left(1 + \frac{1}{2}\right) \times 5890 \times 10^{-10}$$

$$\theta_1 = 31.45^\circ$$



For 5896 A°

$$1.6933 \times 10^{-6} a \sin \theta_{21} = \left(1 + \frac{1}{2}\right) \times 5896 \times 10^{-10}$$

$$\theta_1 = 31.49^\circ$$

This indicates that two spectral lines are diffracted at two different angles. Thus a single slit can separate the colors of the light in the first (or second onwards) secondary maximum.

Example 2.4 Calculate the angular width of the first secondary maximum, when a slit of $1.6933 \mu\text{m}$ is illuminated by sodium source emitting two wavelengths 5890 A° and 5896 A° .

Solution

The first secondary maximum is bounded by first minimum and the second minimum.

The minimum is given by

$$a \sin \theta = m\lambda$$

Thus we have for 5890 A°

$$\theta_1 = \frac{1 \times 5890 \times 10^{-10}}{1.6933 \times 10^{-6}} = 20.26^\circ$$

$$\theta_2 = \frac{2 \times 5890 \times 10^{-10}}{1.6933 \times 10^{-6}} = 44.08^\circ$$

Thus the angular width of the first secondary maximum for 5890 A° is $44.08^\circ - 20.26^\circ = 23.82^\circ$.

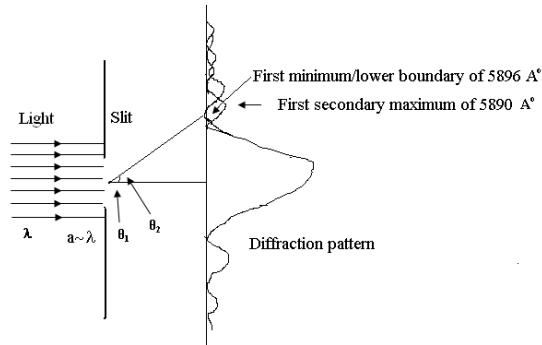
Now for 5896 A°

$$\theta_1 = \frac{1 \times 5896 \times 10^{-10}}{1.6933 \times 10^{-6}} = 20.38^\circ$$

$$\theta_2 = \frac{2 \times 5896 \times 10^{-10}}{1.6933 \times 10^{-6}} = 44.14^\circ$$

Thus the angular width of the first secondary maximum for 5896 A° is $44.14^\circ - 20.38^\circ = 23.76^\circ$

It appears that the width of the first secondary maximum is too broad to be resolved from the next secondary maximum.



Example 2.5 Calculate the dispersive power of a slit having width 1.6933 μm

Solution: From previous problem, we know that this slit diffracts 5890 \AA° at 31.45° and 5896 \AA° at 31.49° . Thus the dispersive power is

$$D.P. = \frac{\theta_2 - \theta_1}{\lambda_2 - \lambda_1} = \frac{41.49^\circ - 31.45^\circ}{5896\text{\AA}^\circ - 5890\text{\AA}^\circ} = 0.0067 \text{ deg/ \AA}^\circ$$

Example 2.6 Suggest a slit which can give maximum dispersive power for the sodium doublet

Solution:

The sodium doublet has wavelengths $\lambda_1 = 5890 \text{ \AA}^\circ$ and $\lambda_2 = 5896 \text{ \AA}^\circ$. The colors are separated in the secondary maximum

$$\text{asin}\theta = \left(m + \frac{1}{2}\right)\lambda$$

Taking $m=1$ $\text{asin}\theta = 1.5\lambda$

Above equation indicates that if $a < 1.5\lambda$, then θ will become indeterminate, and diffraction will not take place. Considering the higher value of λ , we have $1.5 \times 5896 \text{ \AA}^\circ = 8844 \text{ \AA}^\circ$. Thus a slit having width smaller than 8844 \AA° will not diffract the sodium doublet. For this slit, we have

Angle of diffraction for 5890 \AA° , $\theta = \sin^{-1} \left(\frac{1.5 \times 5890}{8844} \right) = 87.42^\circ$ and for 5896 \AA° it is $\theta = \sin^{-1} \left(\frac{1.5 \times 5896}{8844} \right) = 90^\circ$. Thus the dispersive power is

$$D.P. = \frac{90 - 87.42}{5896 - 5890} = 0.43 \text{ deg/ \AA}^\circ$$

This is the maximum possible dispersive power for sodium doublet. Note that the dispersive power depends upon which spectral lines are to be dispersed.

Example 2.7 Calculate the angular width of the first secondary maximum, when a slit having width $1.6933 \mu\text{m}$ is illuminated by sodium doublet having wavelengths 5890 A° and 5896 A°

Solution The first secondary maximum is bounded by first minimum and the second minimum.

The minimum is given by

$$a \sin \theta = m\lambda$$

Thus we have for 5890 A°

$$\theta_1 = \frac{1 \times 5890 \times 10^{-10}}{1.6933 \times 10^{-6}} = 20.26^\circ$$

$$\theta_2 = \frac{2 \times 5890 \times 10^{-10}}{1.6933 \times 10^{-6}} = 44.08^\circ$$

Thus the angular width of the first secondary maximum for 5890 A° is $44.08^\circ - 20.26^\circ = 23.82^\circ$.

Now for 5896 A°

$$\theta_1 = \frac{1 \times 5896 \times 10^{-10}}{1.6933 \times 10^{-6}} = 20.38^\circ$$

$$\theta_2 = \frac{2 \times 5896 \times 10^{-10}}{1.6933 \times 10^{-6}} = 44.14^\circ$$

Thus the angular width of the first secondary maximum for 5896 A° is $44.14^\circ - 20.38^\circ = 23.76^\circ$

2.3 DOUBLE SLIT (OPTIONAL)



The diffraction and interference effects are combined

Let us consider how diffraction pattern changes when we move from single slit to multiple slit. The usual diffraction described by

$$I_\theta = I_m \left(\frac{\sin \alpha}{\alpha} \right)^2, \alpha \text{ has its known meaning} \quad \dots 2.5$$

It may be recalled that the secondary maxima are considerably broad. Now let us see what happens when one slit is added. To avoid the complexity of the discussion, let us keep the width of each slit extremely small. Now what is observed on the screen is the usual double slit interference pattern, described by

$$I_\theta = 4I_m \cos^2 \beta, \text{ where } \beta = \pi \frac{d}{\lambda} \sin^2 \beta, \text{ where } d \text{ is the distance between two slits} \quad \dots 2.6$$

The maxima are sharp enough. This case cannot involve diffraction, as $a \ll \lambda$. Now what will happen if the slit in the latter case are as widened as wavelength of light? This case can be

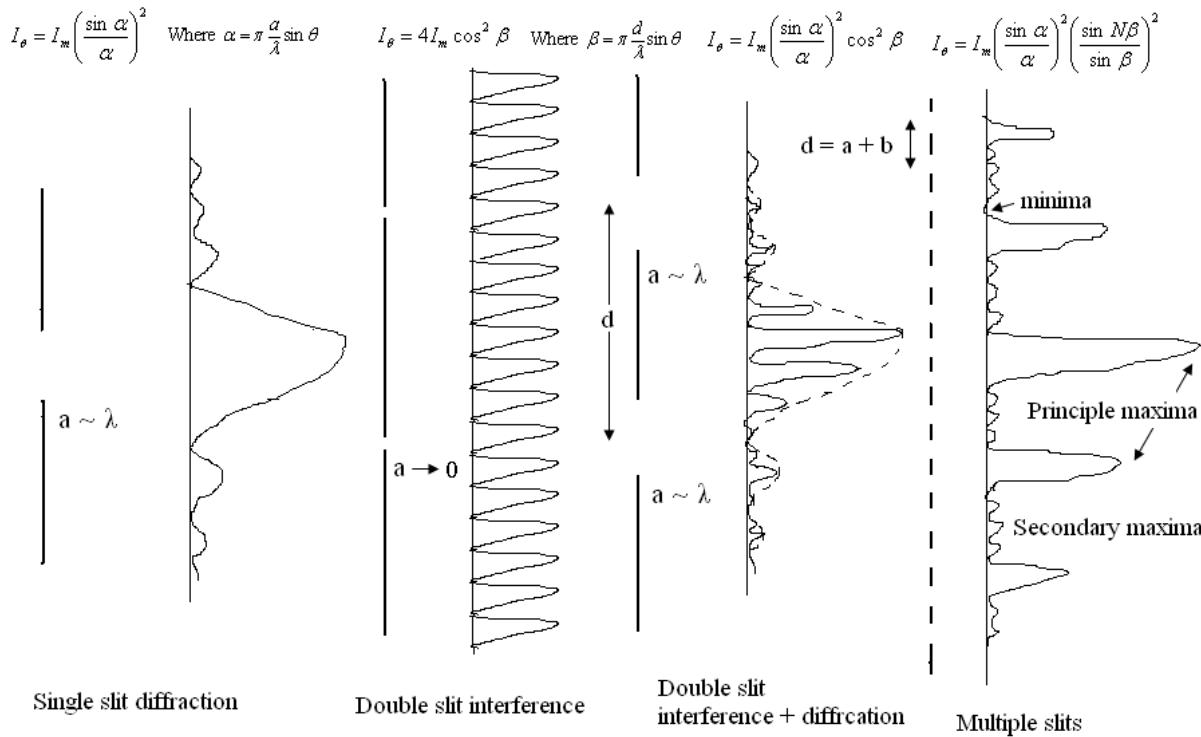


Figure 2.4 Single slit, double slit with extremely narrow slit, the double slit with slit width $a \sim \lambda$ and multiple slits

explained in two ways. As the slits are widened, the diffraction is allowed to take place, which affects the original interference pattern. Thus the resultant intensity in the third case is given by the product of the intensity formulae for the single slit and double slit interference. This formula is

$$I_\theta = 4I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \cos^2 \beta \quad \dots(2.7)$$

Thus the intensity at a given point on the screen in the third figure will be a product of corresponding intensities on the screen in the Fig. a and Fig. b. Thus the set of maxima in the interference pattern multiply with the minima, the resultant intensity decreases. In case of principle maxima in Fig a, the resultant intensity will be enhanced. Further, at the cost of suppressing of the intensities of interference maxima, due to their product with the minima, the intensity of the other interference maxima will increase. It is apparent from eq. 2.7 that it will increase by a factor of 4. This phenomenon can be also explained by considering the interference of diffraction patterns given by eq. 2.5., with a phase difference of β . Somewhere the patterns will interfere constructively and somewhere destructively.

2.4 MULTIPLE SLITS (Optional)

How to produce bright and sharp maxima

Thus diffraction suppresses some maxima in the double slit interference and consequently brightens the others. Further, it is observed that for single slit, there is only one minimum between the principle maximum and the secondary minimum, however, the number of minima and suppressed maxima increases from both the sides of secondary maximum. This narrows its width. Now as the number of slits increases, all these effects enhance, (which are intensification and narrowing of secondary maxima at the cost of increased suppressing due to amplified diffraction effects (due to increase in the number of diffracting slits). Finally when the number of slits, say N increases to a very large value, say 15000-20000, the suppressed maxima become so weak that they become secondary maxima. The formula for the intensity of the diffraction pattern due to multiple slits can be shown to be equal to

$$I_\theta = I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \left(\frac{\sin N \beta}{\sin \beta} \right)^2 \quad (2.8)$$

It may be noted that when N (the total number of slits) is substituted as 2, eq. 2.8 reduces to eq. 2.7, and if N=2 with narrow slits, then with $\alpha \rightarrow 0$, $\left(\frac{\sin \alpha}{\alpha} \right) \rightarrow 1$ and eq. 2.7 reduces to 2.6. If N=1, then we get eq 1.5. Thus it can be said that single slit as well as double slit are the special cases of multiple slits.

At the cost of suppressing, the other maxima (originally the secondary maxima in case of single) are intensified (intensity increases by a factor of N^2 , when the number of slits increases by N). Due to such increased intensities, they are called as principle maxima (of order, m = 0, 1, 2 3 etc.). It will be later shown that when the number of slits increases to N, there occur N-1 minima and N-2 secondary maxima between consecutive maxima. As the space between the principle maxima is occupied by several minima and secondary maxima, the principle maxima become extremely narrow. Thus in brief, multiple slits produce intensified and sharp maxima. In any cases of diffraction, single or multiple slits, angle of diffraction θ is proportional to wavelength λ . Thus different colors are diffracted to different angles owing to different wavelengths. Thus colors are separated and spectrum is formed. However the spectrum of one or limited number of slits is weak, diffused and unresolved and the spectrum of very large number of parallel slits is bright, sharp and therefore well resolved. Later on we will show that large numbers of slits are not advantageous as regards to dispersive power, but they provide large resolving power. Such large number of parallel slits is called **Diffraction Grating** (practically grating means mesh)

Compulsory

2.6 DIFFRICATION GRATING

A super prism, the basis of spectroscopy

Reportedly, Joseph Von Fraunhofer was the first Physicist to fabricate a grating consisting of nearly 200 slits. This was done by winding a thin wire having 200 loops over a screw. However, the technique was soon improved by Rowland by constructing machines known by his names. The machines which have been successively improved are the ones which are attached with a fine diamond point. With extremely controlled motion of this point, equidistant lines are grooved on plain glass plates. The scratched parts become opaque to the light, while the space between the scratched acts like slits. The typical numbers of slits in the gratings are in the range 15000-20000 per inch. A typical sketch of the diffraction grating and its intensity pattern is shown in Fig. 2.4. The quality of diffraction grating depends upon how small its grating element is ($d = a + b$, where a is the width of slit (transparency) and b is the width of opacity), how many slits (N) and how many slits per unit length (N') it has.

By applying phasor method to calculate the resultant intensity at point having an angle of diffraction θ , it can be shown that

Eqns (2.9 to 2.12) are compulsory but their derivations re optional



$$I_{\theta m} = \lim_{\beta \rightarrow m\pi} I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \left(\frac{\sin N\beta}{\sin \beta} \right)^2 \quad (2.9)$$

Eqn 2.9 describes the intensity at all the points, however, a few typical intensities such as principle maxima, minima can be described.

In general, the factor in the first bracket of eqn 2.9 corresponds to the contribution of the single slit, while the second bracket, which appears due to superposition of the diffraction patterns of N number of slits, is more significant. Thus for principle maxima, $\sin \beta$ needs to be tend to zero. This requires $\beta \rightarrow m\pi$. Thus

$$I_{\theta m} = \lim_{\beta \rightarrow m\pi} I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \left(\frac{\sin N\beta}{\sin \beta} \right)^2$$

However when $\beta \rightarrow m\pi$, $N\beta \rightarrow mN\pi$ and thus $\sin \beta$ as well as $\sin N\beta \rightarrow 0$ at the same time. This is indeterminacy and it can be resolved by applying L' Hospital rule (See appendix to chapter 2). This requires differentiation of the numerator and the denominator, independently. Thus

$$I_{\theta m} = I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \lim_{\beta \rightarrow m\pi} \left(\frac{\frac{d}{d\beta} \sin N\beta}{\frac{d}{d\beta} \sin \beta} \right)^2$$

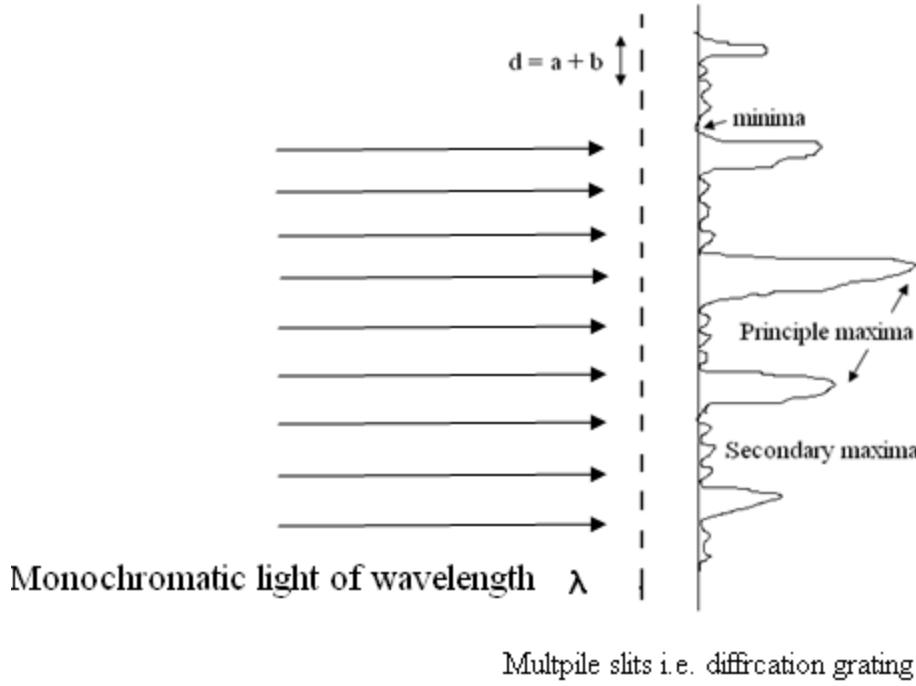


Figure 2.4. Diffraction grating and its intensity pattern

$$I_{\theta m} = I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \lim_{\beta \rightarrow m\pi} \left(\frac{N \cos N\beta}{\cos \beta} \right)^2$$

$$I_{\theta m} = N^2 I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \quad \dots(2.10)$$

Thus the intensity of principle maxima is N^2 times greater than the corresponding intensity of secondary maxima in the single slit. The condition for the maxima to occur is thus $\beta = m\pi$. Solving this condition,

$$\beta = m\pi$$

$$\frac{\pi}{\lambda} d \sin \theta = m\pi$$

$$d \sin \theta = m\lambda \quad \dots(2.11)$$

Eqn 2.11 predicts, at what θ, m^{th} principle maxima occurs for a grating having grating element d . This equation is called as condition of maxima

For minima $\sin\beta$ needs to be equal to zero, thus

$$\beta = m'\pi$$

By substituting β , we get

$$d \sin \theta = \frac{m'}{N} \lambda \quad \dots (2.12)$$

Where m' is an integer designating the sequence number of minima. m' is prohibited to take the values given by mN , as then the equation 2.12 reduces in the condition for the maxima

All these problems are compulsory



Example 2.8 A diffraction grating having 15000 slits per inch is exposed to a Krypton source emitting wavelengths 4000 \AA for violet, 4500 \AA for blue, 5000 \AA for green, 5790 \AA for yellow, 6000 \AA for orange, 6500 \AA for red and 7000 \AA for brown. Calculate the angle of diffraction of these lines in the first order.



Solution:

The condition for the maxima of the grating is

$$ds \sin \theta = m \lambda$$

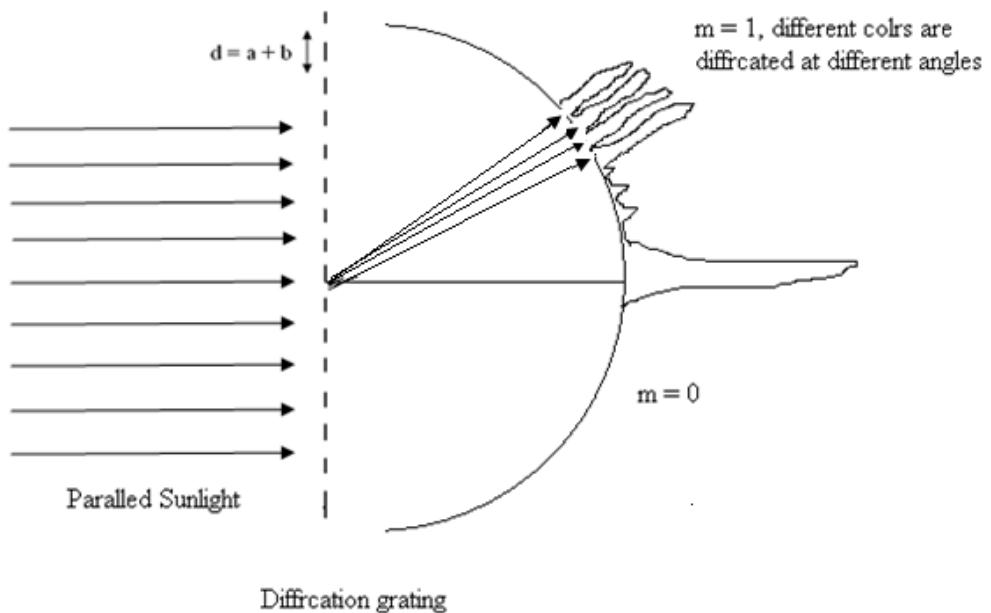
The grating has 15000 slits in 1 inch

Thus it has 15000 grating elements (d) in 1 inch

$$\text{Thus 1 grating element } d = \frac{1''}{15000} = \frac{2.54}{15000} \text{ cm} = \frac{2.54 \times 10^{-8}}{15000} \text{ A}^\circ = 16933.33 \text{ A}^\circ$$

Considering first order i.e. $m = 1$

$$\theta = \sin^{-1} \frac{\lambda}{d}$$



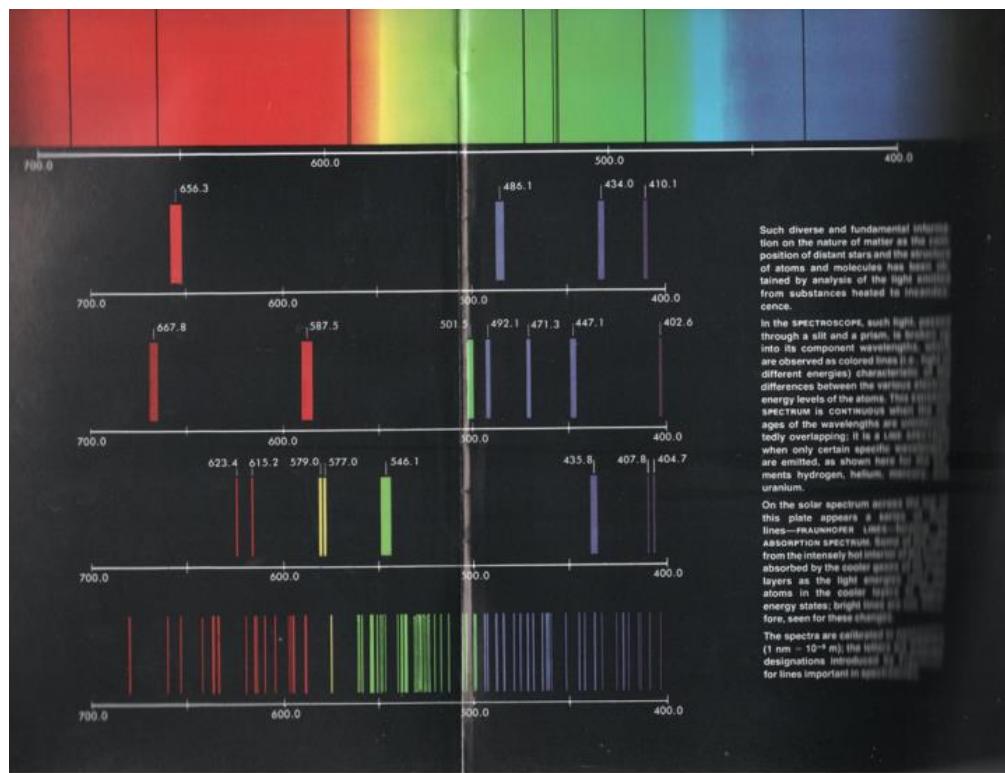
For violet

$$\theta = \sin^{-1} \frac{4000}{16933.33} = 13.66^\circ$$

For blue

$$\theta = \sin^{-1} \frac{4500}{16933.33} = 15.41^\circ$$

Similar calculations for green, yellow, orange, red and brown give $\theta = 17.77^\circ, 19.67^\circ, 20.75^\circ, 22.57^\circ$ and 24.42° respectively. This clearly indicates that grating can separate the colors of any source and form its spectrum. The analysis of such spectra leads to understanding of atomic and molecular structures of the elements and compounds. The effects such as Raman effect, Zeeman Effect can also be studied using the grating. Thus gratings have become inevitable source in spectroscopy. It may be noted that each element and each compound has its characteristic spectrum which just acts like a fingerprint of that element. This makes it possible to identify the species by analyzing its spectrum. The fact that Sun contains hydrogen and helium has also been established by spectral analysis.



**The spectrum of any element or compound can be obtained by using diffraction grating.
The analysis of such spectra is the basis of spectroscopy**



The colors on CDs are due to diffraction. Indeed, a CD is itself is a diffraction grating

Example 2.9: Calculate the angular separation between the sodium doublet (5890 \AA° and 5896 \AA°) by using a grating having 15000 slits per inch in the first second and the third order.

Solution:

With $m=1$ and $16933.33 \text{ \AA}^{\circ}$, the condition for the maxima becomes

$$\theta = \sin^{-1} \frac{5890}{16933.33} \text{ and } \theta = \sin^{-1} \frac{5896}{16933.33}, \text{ which gives } \theta_1 = 20.36^{\circ} \text{ and } \theta_2 = 20.38^{\circ}$$

With $m=2$, we have

$$\theta = \sin^{-1} \frac{2 \times 5890}{16933.33} \text{ and } \theta = \sin^{-1} \frac{2 \times 5896}{16933.33}, \text{ which gives } \theta_1 = 44.08^{\circ} \text{ and } \theta_2 = 44.14^{\circ}$$

With $m=3$, we have

$\theta = \sin^{-1} \frac{3 \times 5890}{16933.33}$ and $\theta = \sin^{-1} \frac{3 \times 5896}{16933.33}$, which gives θ_1 and θ_2 to be indeterminate. In fact a transmission grating can produce the spectra only up to 90° . For the first two spectra the angular separation with the sodium doublet is 0.02° for the first order and 0.06° . This means that grating gives higher angular separation in higher order. However, the higher orders give weak intensities, thus first order is preferred.

Example 2.10 Calculate the angular width of the first order maxima, if a grating having 15000 slits per inch is exposed to a laser having wavelength 6328 \AA° . Repeat the calculations for a grating having 20000 slits per inch and record your conclusions.

Solution:

We know that the 15000 minimum is absent as there 1st maximum exists. Thus the first maximum is bounded by 14999th and 15001th minima.

The condition for the minima for the grating is

$$dsin\theta = \frac{m'}{N}\lambda$$

Thus for the above mentioned minima

$$dsin\theta = \frac{14999}{15000} 6328 = 21.94^{\circ}.$$

And

$$dsin\theta = \frac{15001}{15000} 6328 = 21.95^\circ.$$

Thus the angular width for the first principle maximum is 0.01° . In the Ex 2.4, the width of the first secondary maximum in case of the single slit came out to be 23.82° . This is too large to resolve the spectral lines, however, when the number of the slits are increased from 1 to 15000, the secondary maximum intensifies by 15000^2 and narrows down to 0.01° . This is the advantage of multiple slits. It is not necessary to repeat the calculations for the 20000 slits. The maximum intensifies by 20000^2 and narrows further.

Example 2.6: The first principle maximum of the grating having grating element 16933.33° illuminated by laser having wavelength 6328A° occurs at 21.94° . Assume that the width of the slit is 8477 A° What is its relative intensity? Interpret the results

Solution:

The relative intensity of the principle maximum is given by

$$\frac{I_\theta}{I_m} = N^2 \left(\frac{\sin\alpha}{\alpha} \right)^2, \text{ where } \alpha = \pi \frac{a}{\lambda} \sin\theta$$

We now calculate α ,

$$\alpha = \pi \frac{8477}{6328} \sin 21.94 = 1.57^\circ$$

Thus

$$\frac{I_\theta}{I_m} = N^2 \left(\frac{\sin\alpha}{\alpha} \right)^2 = 15000^2 \left(\frac{\sin 1.57}{1.57 \times \frac{\pi}{180}} \right)^2 = 225000000 \times \left(\frac{0.027}{0.027} \right)^2 = 225000000$$

This result will appear perplexing, if I_m is considered as the intensity of the principle maximum of the grating. This will indicate that the intensity of the first principle maximum is 225000000 times greater than the central maximum. This not possible, as the intensity decreases with the order of the maxima. The only possibility which makes the thing unambiguous is to take I_m as the intensity of the central maximum of the single slit. This so because after taking the limit $\beta \rightarrow m\pi$, the term $\left(\frac{\sin N\beta}{\sin \beta} \right)^2$ maximizes to N^2 . Thus the term I_m in the formula $I_\theta = I_m \left(\frac{\sin\alpha}{\alpha} \right)^2 \left(\frac{\sin N\beta}{\sin \beta} \right)^2$ corresponds to the intensity of the central maximum of the single slit, while I_m in the formula $I_\theta = N^2 I_m \left(\frac{\sin\alpha}{\alpha} \right)^2$ seems to be corresponding to the intensity of central maximum of the single slit. Without this consideration, the result remains unexplained.

Eqns. (2.13), (2.14) and (2.18) are compulsory, but their derivations are optional



2.8 DISPERSIVE POWER AND RESOLVING POWER OF A GRATING

A grating having good dispersive power may not have good resolving power and ...vice versa

Consider the grating formula

$$dsin\theta = m\lambda$$

As, for the given grating and given order, d and m are constant, we understand that $\theta \propto \lambda$. This means that grating separates the colors of light. But how effectively? This is decided by the Dispersive Power (D.P.) and Resolving Power (R.P.). Dispersive power relates to the ability of the grating to produce a significant and largest possible angular separation between the spectral lines. However, or any grating of any dispersive power, when the difference between the spectral lines tends to zero, the spectral images of the lines will overlap. In case of excessive overlapping, the resolution is lost. In such cases, if the dispersive power cannot be increased beyond a limit, then in overlapping condition only, the spectral images can be made sharp by increasing the total number of the slits in the grating. As the spectral images become narrow and sharp, the region of overlapping decreases, leading to a possibility of resolution.

Differentiating the grating formula

$$dcos\theta d\theta = md\lambda$$

$$\frac{d\theta}{d\lambda} = \frac{m}{dcos\theta} \quad \dots(2.13)$$

The quantity on the L.H.S. i.e. $\frac{d\theta}{d\lambda}$ is called as dispersive power i.e. dispersion per unit wavelength difference. Thus if a grating produces higher angular separation between the spectral lines having smaller difference, then the grating has better dispersive power. It is apparent from eqn (2.13) that dispersive power increases with the order of the spectrum, decrease in the grating element d , and decreases in $cos\theta$, where θ is the angle of the diffraction of the first of the two spectral lines, which are dispersed. We can increase the dispersive power by increasing the order of the spectrum i.e. m , however, then the intensity of the spectrum will decrease. Thus lowest value of m is desirable. Thus we choose $m = 1$. For lower orders θ is also small, thus $cos\theta$ can be taken approximately as 1. Further, as the grating containing N' number of slit per unit length will contain N' number of grating elements per unit length, we take $d = \frac{1}{N'}$. By considering this, the formula for the dispersive power, 2.13, becomes

$$\frac{d\theta}{d\lambda} = mN' \quad \dots(2.14)$$

Thus it appears that dispersive power can be increased indefinitely by increasing the slit density as much as it is required. This is impossible theoretically as well as practically, as when N' is increased d will decrease and somewhere it will fall below the lowest possible wavelength of light. Now the grating equation ($ds\sin\theta = m\lambda$) indicates that θ will become indeterminate even for the lowest significant value of $m = 1$, if $d < \lambda$

Example 2.7 Calculate the dispersive power of the grating containing 15000 slits per inch and 20000 slits per inch

Solution:

We have

$$N' = 15000 \text{ slits per inch} = \frac{15000}{2.54 \times 10^8} \text{ slits per } A^\circ = 5.9 \times 10^{-5} \text{ slits per } A^\circ$$

$$\text{Thus } d = \frac{20000}{2.54 \times 10^8} \text{ slits per } A^\circ = 7.87 \times 10^{-5} \text{ slits per } A^\circ$$

$$\text{We have } D.P. = mN' = 1 \times 5.9 \times 10^{-5} \text{ rad}/A^\circ = 5.9 \times 10^{-5} \text{ rad}/A^\circ \times \frac{180}{\pi} \text{ deg/rad}$$

$$D.P. = mN' = 1 \times 5.9 \times 10^{-5} \text{ rad}/A^\circ = 5.9 \times 10^{-5} \text{ rad}/A^\circ \times \frac{180}{\pi} \text{ deg/rad} = 0.034 \text{ deg}/A^\circ$$

$$D.P. = 0.034 \text{ deg}/A^\circ = 2.04 \text{ min}/A^\circ$$

Similar calculations for 20000 slits per inch lead to $D.P. = 16.25 \text{ min}/A^\circ$

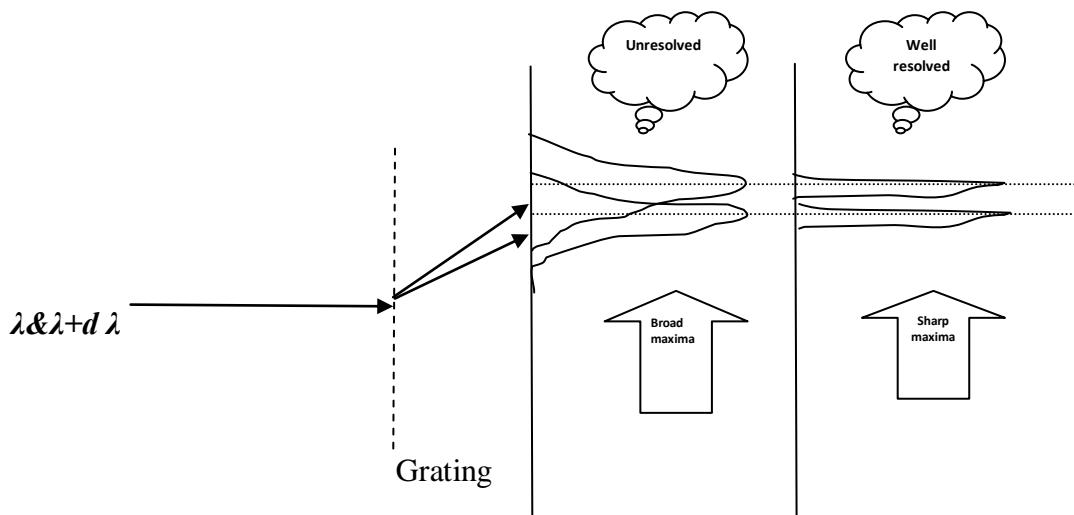


Figure 2.5 A grating of maximum dispersive power two spectral lines only if it makes them sharp

Let us now consider that, a grating having maximum possible dispersive power is exposed to two spectral lines having wavelengths λ and $\lambda+d\lambda$. Let $d\lambda$ be extremely small. The grating will diffract the lines at, say θ and $\theta + d\theta$, but $d\theta$ will be very small, so small that the spectral lines will overlap on each other (refer Fig. 2.5). To avoid overlapping, what can only be done is to

make the lines sharp, as the dispersive power is at its maximum limit, and the angular separation between the spectral lines cannot be



Lord Rayleigh 1842-1919. The Nobel Prize in Physics 1904 was awarded to Lord Rayleigh "for his investigations of the densities of the most important gases and for his discovery of argon in connection with these studies". Along with his several contributions spanning almost every area of Physics, such as sound, wave theory, color vision, electrodynamics, electromagnetism, light scattering, flow of liquids, hydrodynamics, density of gases, viscosity, capillarity, elasticity, and photography, leading to publication of 446 research papers, one was to specify a criterion of resolution for an optical instrument, known due to his name.

increased. The spectral lines can be made sharp, just by increasing the total number of slits, irrespective of the space in which they have been compressed. Thus the total number of slits can be increased by keeping slit density constant. Thus if a grating with 20000 slits per inch cannot be improved to 40000 slits per inch then a two inch grating having total 40000 slits can be fabricated. Making a grating with large number of slits increases the sharpness of the spectral lines and thus helps them resolve, especially when $d\lambda$ is extremely small. This relates to another quality of grating called Resolving Power (R.P.). Indeed what is to be proven here is that while the dispersive power depends upon the slit density, the resolving power depends upon total number of slits.

For a grating of any dispersive power, there is certain minimum $d\lambda$, for which the angular separation is so small that the spectral images overlap on each other. If the overlapping is excessive, then the resolution is not possible, and if the angular separation increases then at certain state, in spite of overlapping the separation just becomes sufficient to barely resolve the spectral images. This relates to Rayleigh's criterion of resolution as explained in Figure 2.6. The part b of the Fig () indicates that for bare resolution of the spectral lines, the central maximum of the diffraction pattern of the one line needs to overlap on the first minimum of the other and vice versa. In this case the composite intensity due to overlapped diffraction patterns in the region of overlapping falls slightly below the maximum intensity of the two images. If the overlapping exceeds this limit then the summed up intensity in the overlapped region exceeds the individual intensities and thus the depression in the center required for the resolution disappears and we observe a single image indicating only one object even though there are two objects (say stars)

Rayleigh's criterion, though a criterion without any proof, remains an empirical judgment that defines the allowed extent of overlapping of the diffraction patterns of the two objects, and it has worked successfully over last few decades, for all the instruments like cameras, binoculars, telescopes, microscopes, where two closed images and their overlapped diffraction patterns are involved.

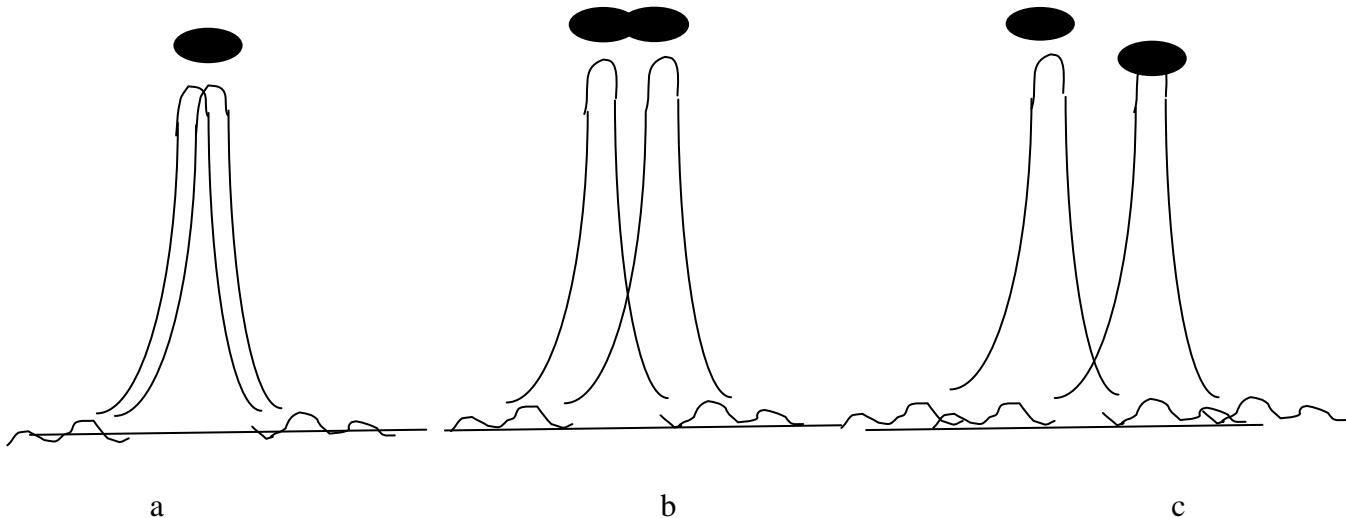


Figure 2.6 Rayleigh's criterion of resolution a. unresolved b. barely resolved c. well resolved

2.8 RESOLVING POWER OF THE GRATING

Why gratings are made up of enormously large number of slits

Consider a diffraction grating having N' number of slits per unit length, N number of total slits and grating element d . Let the grating be exposed to two spectral lines having wavelengths λ and $\lambda+d\lambda$. The spectral lines will be diffracted at two different angles, say θ and θ' . The angular separation between the lines will be zero in the zeroth order, as all the gratings have zero dispersive power in the zeroth order (eqn (2.14)). However, as per the equation 2.13 the angular separation between the images will gradually increase with the order. For lower order, the angular separation will be smaller, and overlapping will be more. The net intensity in the overlapped region will be higher than the individual intensities of the λ and $\lambda+d\lambda$. Thus resolution will be lost. However, as the order of the spectral lines increases, the angular separation will increase and the region of overlapping will decrease. During this process, the net intensity in the overlapped region will become smaller than the individual intensities, and a small but noticeable depression will appear between the spectral images of the λ and $\lambda+d\lambda$. It is this depression which will help the observer to resolve the spectral lines, though barely. When such depression occurs, the Rayleigh's criterion is satisfied. For gratings Rayleigh's criterion cannot be satisfied at central maximum and first minimum, as for any grating the central maximum of any spectral lines the central maximum should fall on each other. Thus for gratings, two spectral lines are barely resolved, when the m^{th} maximum of lower wavelength (λ) overlaps on the $(mN-1)^{\text{th}}$ minimum of higher wavelength ($\lambda+d\lambda$) and vice versa, i.e. the m^{th} principle maximum of higher wavelength $\lambda+d\lambda$ overlaps on the $(mN+1)^{\text{th}}$ minimum of λ . Thus the Rayleigh's criterion is satisfied at, θ and also at θ' . Note that for λ as well as $\lambda+d\lambda$, the mN^{th} minimum is forbidden, as this position is reserved for m^{th} maximum. Thus for any spectral line, the minimum just before

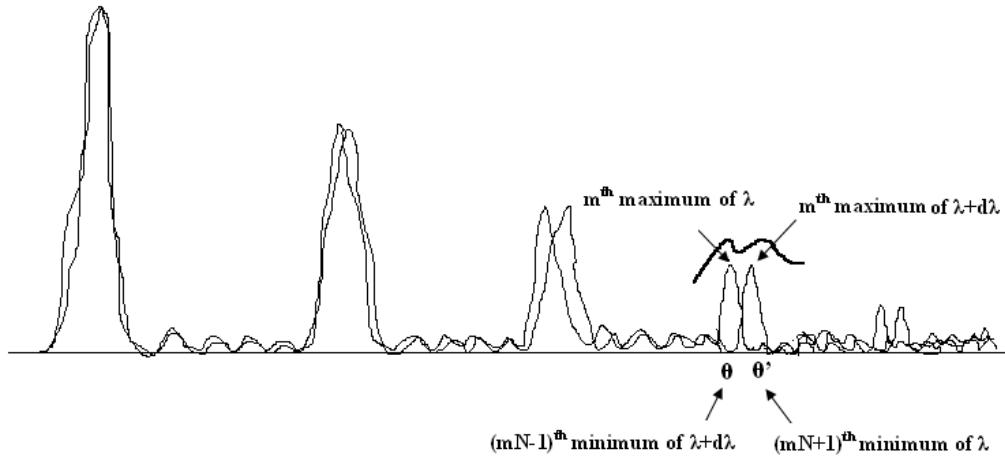


Figure 2.7 For a grating, Rayleigh's criterion is not satisfied in the zeroth order, but in mth order

the m^{th} maximum is $(mN-1)^{th}$ -minimum and the minimum just next to m^{th} maximum will be $(mN+1)^{th}$ minimum

We know that the conditions of maxima and minima are given by

$$dsin\theta = m\lambda \quad \dots(2.15)$$

$$dsin\theta = \frac{m'}{N} \lambda \quad \dots(2.16)$$

Applying these conditions at θ , we get

$$dsin\theta = m\lambda$$

$$dsin\theta = \left(\frac{mN - 1}{N}\right)(\lambda + d\lambda)$$

The above two eqns are simultaneous eqns, as Rayleigh's criterion is satisfied here. As L.H. S. are same, we equate the R.H.S.

$$m\lambda = \left(\frac{(mN - 1)}{N} \right) (\lambda + d\lambda)$$

$$\Rightarrow mN\lambda = mN\lambda + mNd\lambda - \lambda - d\lambda$$

$$\Rightarrow \frac{\lambda + d\lambda}{d\lambda} = mN$$

$$Resolving\ power\ of\ a\ grating = R.P. = \frac{\lambda + d\lambda}{d\lambda} = mN \quad \dots(2.17)$$

In the above equation, $d\lambda$ represents the limit of resolution, as when the Rayleigh's criterion is satisfied, $d\lambda$ has smallest permissible value.

We can also derive a formula for R.P. by using the Vice Versa of the Rayleigh's criterion of resolution. Accordingly, now, the m^{th} principle maximum of higher wavelength $\lambda + d\lambda$ overlaps on the $(mN+1)^{th}$ minimum of lower wavelength λ . According to the Fig, this overlapping occurs at θ' .

We now frame the equations () and () for principle maxima and minima as per the above Rayleigh's criterion.

$$dsin\theta' = m(\lambda + d\lambda) \text{ and}$$

$$dsin\theta' = \left(\frac{mN - 1}{N} \right) \lambda$$

We now solve above two eqns simultaneously. *LHS* of both the eqns being identical, we equate the *RHS*

$$m(\lambda + d\lambda) = \left(\frac{mN - 1}{N} \right) \lambda$$

$$\Rightarrow mN(\lambda + d\lambda) = (mN - 1)\lambda$$

$$\Rightarrow mN\lambda + mNd\lambda = mN\lambda - \lambda, \quad \text{rearranging, we get}$$

$$\Rightarrow Resolving\ power\ of\ grating = R.P. = \frac{\lambda}{d\lambda} = mN \quad \dots(2.18)$$

Note that we have defined the resolving power of the grating as either $\frac{\lambda+d\lambda}{d\lambda}$ or $\frac{\lambda}{d\lambda}$. Note that we want to resolve λ and $\lambda + d\lambda$ where $d\lambda$ is the limit of resolution (the smallest difference that can be just resolved). Clearly, smaller the $d\lambda$ better is the resolving power of the grating. This is why $d\lambda$ occurs in the denominator of these formulae. The numerators are either λ or $\lambda + d\lambda$, which are taken as a reference.

Example (2.8) How many slits are required by the grating to barely resolve the sodium doublet in the first order?

Solution:

We have

$$R.P. = \frac{\lambda}{d\lambda} = mN$$

$$N = \frac{5890}{6} = 982 \text{ slits}$$

It appears that for just resolving sodium doublet only 982 slits are sufficient, however, a grating having 15000 LPI will not only resolve them, but will also produce a good angular separation between them.

Example (2.9) What is the limit of resolution of a grating having 15000 LPI in first order? Use the reference wavelength as the average wavelength of light, 5500 A°.

Solution:

$$R.P. = \frac{\lambda}{d\lambda} = mN$$

$$d\lambda = \frac{\lambda}{mN}$$

$$d\lambda = \frac{5500}{1 \times 15000}$$

$$d\lambda = 0.37 \text{ A}^\circ$$

Diffraction due to circular aperture

Refer the photograph showing the diffraction pattern due to circular aperture. We can judge this diffraction pattern, by imagining that the diffraction pattern of a single slit is rotated through

360°. The linear fringes in the single slit diffraction pattern will be converted in to cones and rings. The angular position of the first minima in a single slit is given by

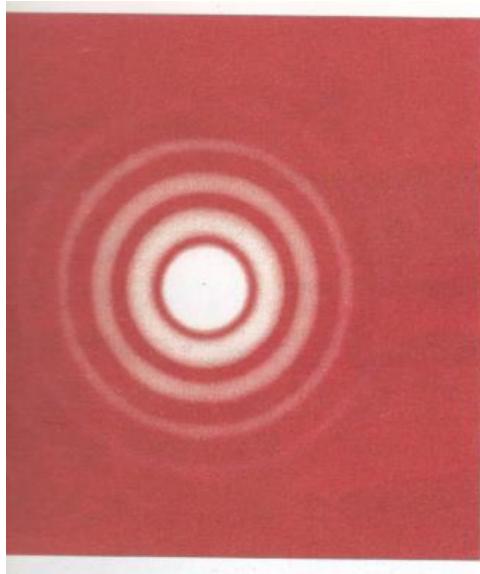
$$a \sin \theta = m\lambda \text{ with } m = 1$$

$$a \sin \theta = 1\lambda$$

In the equation for the first minima of a diffraction pattern due to circular aperture ‘1’ in the above equation is replaced by 1.22. Thus the equation becomes

$$a \sin \theta = 1.22\lambda \dots \text{(first minima in diffraction pattern of a circular aperture).}$$

We know that the first minimum is the boundary of central maxima in the diffraction pattern. Thus the boundary of the central disk (called Airy's disk named due to George Biddell Airy) is also first minimum (designated by)



Diffraction pattern of a circular aperture. Where is the Airy's Disc?

2.9 HUMAN EYE, TELESCOPES, MICROSCOPES, CAMERA ETC.

Resolving power is governed by diffraction effects

We have learnt two facts; one is that when the size of the obstacle or the ratio $\frac{a}{\lambda}$ widens, the diffraction effect becomes gradually weak and thus the central maximum becomes narrower. This means that if a telescope of large lens is chosen, the diffracted images of the stars will be narrow, and in such cases when the angle subtended by the stars at the telescope (due to extremely large distance between them and the telescope) becomes extremely narrow the images overlap, but their narrowness allows resolution. This means that bigger telescopes are better.

Whiles, the biggest telescope in the world has 2.00 m lens, making the lens still bigger is difficult. In such cases, the multi-array telescopes give high resolution. Here the concept of grating plays a role. We have seen that, larger the slits, brighter and sharper are the images. Multi-array telescopes thus produce sharp and bright images of the stars. By applying Rayleigh's criterion of resolution, it can be shown for telescopes and telescope-like instruments that

$$R.P. = \frac{1}{\theta_R} = \frac{d}{1.22\lambda} \quad \dots(2.19)$$

For microscopes, taking large value of **d** i.e. the bigger lens is not viable, thus the wavelength is reduced. Electrons which behave like waves have extremely small; say 1 to 10 lakh times smaller wavelengths as compared to light. Thus electron microscope offers incredible resolution and magnification as compared to optical one. Electron microscope and its adherent Scanning Tunneling Microscope have made nanotechnology possible, nanotechnology would have been beyond the reach of conventional optical microscopes.

Radars with bigger antennas and multi-antenna radars are also based on the fact that a precise control on diffraction effects leads to better resolution. Two human eyes, two human ears are the other examples. Seeing the distant objects with only one eye and listening multi-frequency music with only one ear places a limit on resolution! Why only one mouth then? Probably GOD desires that we should talk less and listen more!



Resolving power of all telescopes such as GMRT or multi-array telescope is governed by diffraction. Why several antennas instead of one?

RAPID REVIEW

Diffraction is not just the bending of light. The obstacles like slit and multi-slit systems produce well analyzable diffraction patterns. Diffraction leads to two foremost applications in optical engineering, one is diffraction grating...a super-prism, which separates the colors of light more effectively than prism and another resolving power of several instruments. The concept of Resolving power is based in Rayleigh's criterion of resolution, which also decided the resolving powers of human eye, telescope, microscope, binocular, camera etc. Multi-array antennas and multi-array Radars are the extensions of the multiple concepts that help for better resolution.

The intensity formula for the single slit diffraction is given by

$$I_\theta = I_m \left(\frac{\sin \alpha}{\alpha} \right)^2$$

The conditions of minima and secondary maxima that follow from the above formula are ,

$$\alpha \sin \theta = n\lambda \text{ and } \alpha \sin \theta = \left(n + \frac{1}{2} \right) \lambda$$

The equation for secondary maxima indicates that diffraction can separate the colors as of light as $\theta \propto \lambda$. However single slit method is a crude method to form the spectrum of light as the secondary maxima are weak in the intensity and broad in width.

When number of slits are increased the maxima are brightened and sharpened. Such system is called diffraction grating, the basic intensity equation of which is given by

$$I_\theta = I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \left(\frac{\sin N\beta}{\sin \beta} \right)^2$$

The principle maxima can be interpreted by taking limit $\beta \rightarrow m\pi$, where we get

$$I_{\theta \max} = N^2 I_m \left(\frac{\sin \alpha}{\alpha} \right)^2 \text{ and } d \sin \theta = m\lambda \text{ (grating formula)}$$

Grating is a dispersive device, its dispersive power is given by

$$\frac{d\theta}{d\lambda} = \frac{m}{d \cos \theta} \approx mN'$$

No grating can have unlimited dispersive power, when the maxima for the grating having highest dispersive power overlap, the resolving power can be increased by making the maxima sharp. Rayleigh's criterion of resolution leads to

$$R.P. = \frac{\lambda}{d\lambda} = \frac{\lambda + d\lambda}{d\lambda} = mN$$

It may be noted that the dispersive power depends upon slit density and the resolving power on the total number of slits. Principally, both these parameters are independent of each other.

APPENDIX I TO CHAPTER 2

Phasor and Phasor diagrams

Consider Fig on the next page. A wave can be represented by a vector. While doing third three considerations are to be followed. The length of the vector should be proportional to the amplitude of the wave and the angle subtended by the vector w.r.t. X axis should be equal to the instantaneous phase of the vector. When this is followed, the y projection of the vector becomes the instantaneous displacement of the wave. Such vectors are called as phasors as their orientation in the space depends upon the phase of the wave.

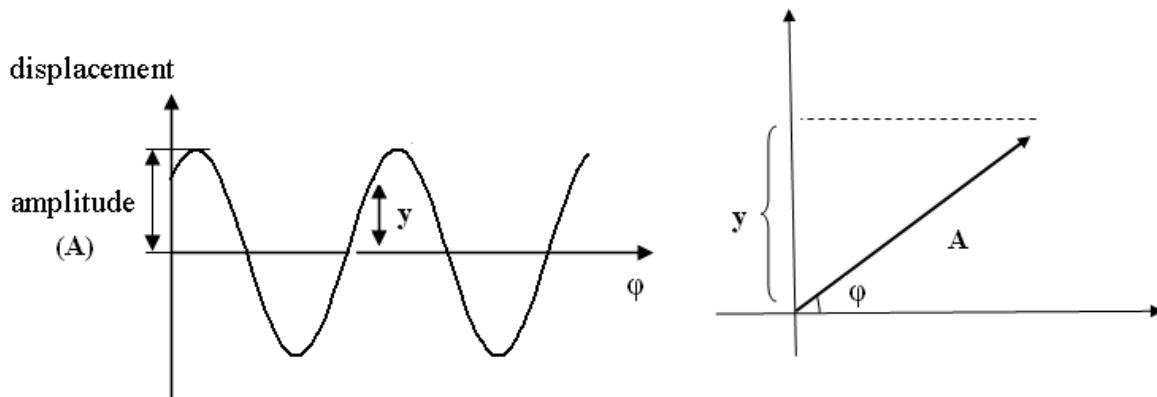


Figure (2.8) Representing a wave by a vector

Consider following Fig. In the first part two waves are superimposing on each other. The instantaneous phase of the first and the second wave are ϕ_1 and ϕ_2 respectively. The phased difference is $\Delta\phi$. These two waves can be added with the help of phasor diagrams. The first wave is represented by a horizontal phasor. The second is represented by a phasor whose bottom joins with the tip of the first phasor. The angle between the two phasors is the phase difference between the corresponding waves. The resultant can be obtained by joining the bottom of the first phasor with the tip of the second phasor. The length of resultant phasor gives the amplitude of the resultant wave. When this amplitude is squared, the resultant intensity is obtained. This method can be extended for three or more than three waves also. Note that in the later part of the Fig, the angle between the first phasor extended in the forward direction and the third phasor extended in backward direction is the sum of the angles between the first and second and second and third phasor. When this method is extended to N number of superimposing waves being represented by N number of phasors in the phasor diagrams, the angle between the first phasor extended in the forward direction and the Nth vector extended in the backward direction is the summation of the individual angles between the neighboring phasors. In case of large number of phasors, the phasor diagram appears like an arc rather than a polygon.

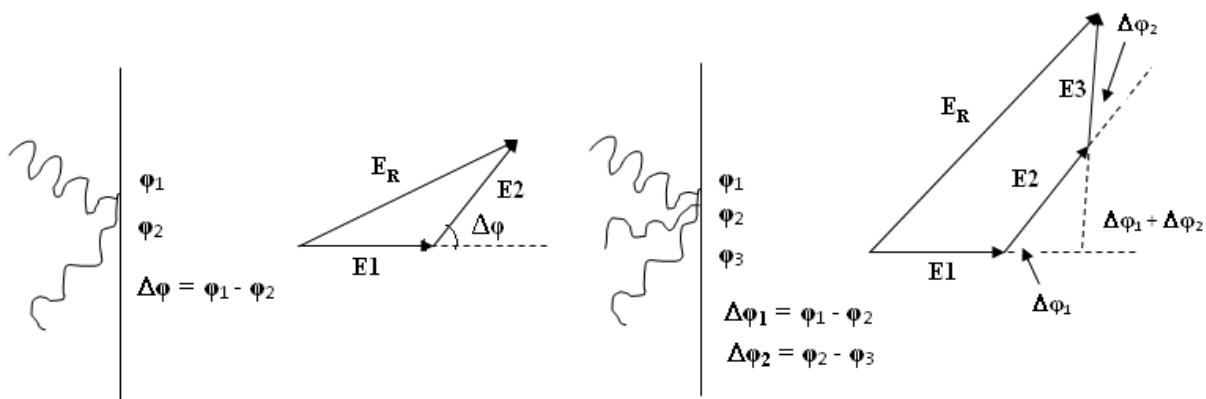


Figure (2.9)Adding Phasors: Phasor diagram

APPENDIX II TO CHAPTER 2

Diffraction grating and Spectroscopy

Diffraction grating is a dispersive device with incredible dispersive power and resolving power. Interestingly the elements in the periodic table represent their unique spectra without any duplication, just like a fingerprint or a signature. Thus an element or even a molecular species can be identified by analyzing its spectrum. This requires diffraction grating. The spectrum of an element or compound is related to its atomic or molecular structure. Indeed the only experimental way to arrive at the model of an atom or molecule is through spectroscopy itself. Diffraction grating is also required to record the minute shift in the wavelengths that occurs in Raman effect (due to scattering) or Zeeman effect (due to effect of magnetic field on the electronic orbital)

EXERCISES

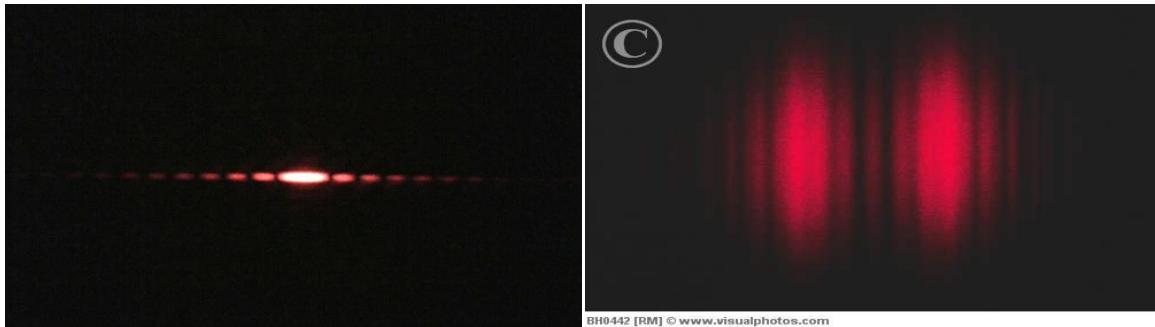
In questions of science, the authority of a thousand is not worth the humble reasoning of a single individual.

Galileo Galilei

General

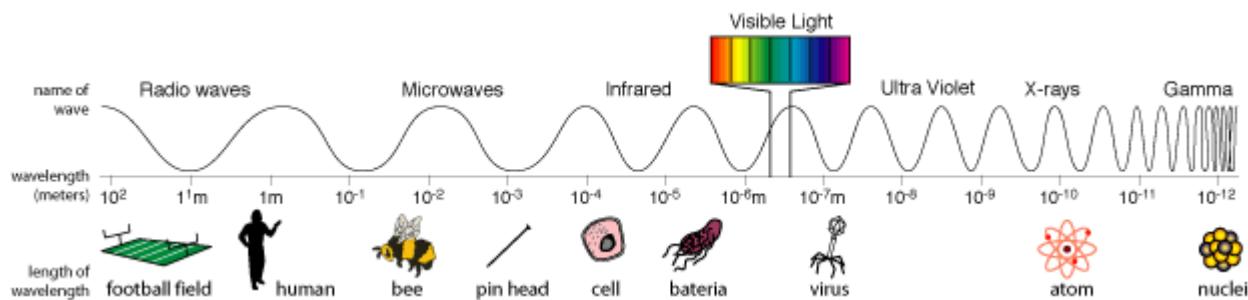
1. Consider any optical instrument of your choice, say a Human eye and indicate how diffraction is involved in it.
2. Mention any one application of diffraction in optical engineering or in daily life
3. Diffraction is of just the bending of light, then what else it is?

4. Diffraction is an extension of interference and interference, s special case of diffraction. How?
5. Consider a man standing behind the door and speaking. You can listen him but can't him. Why?
6. Idealized diffraction is impossible. How?
7. Idealized interference is impossible. Why?
8. Consider a razor blade exposed to light, the image is not sharp, why?
9. Can diffraction be explained using particle character of light? Why? Why not?
10. When the obstacle is extremely bigger than light the diffraction weakens. Can this be explained by using Huygens theory of secondary wavelets? How?
11. Diffraction is there due to secondary wavelets emanating from various points in the obstacle. What then happens to diffraction and the secondary wavelets when the obstacle is removed?
12. It seems that, diffraction is possible only when the secondary wavelets emanate from the obstacles? Why diffraction disappears when the obstacle is removed even if Huygen's secondary wavelets still exists?
13. Consider a slit having width $a = \lambda$, $a = 5\lambda$ and $a = 20\lambda$ and calculate the angular width of the central maximum. Indeed the results indicate that diffraction becomes weak, when a/λ ratio is increased. How?
14. In single slit diffraction, when the wave front passes through the slit, the points in the slit become source of secondary wavelets and produce diffraction effects as shown below. Now consider a thin wire exposed to light, in this case also the light will be obstructed a produce diffraction effects. From where do the secondary wavelets arise then?
15. A CD satisfies all conditions for a reflection type of grating. Do you propose that instead of purchasing a costly grating, a cheap CD itself can be used as a grating in all the applications?

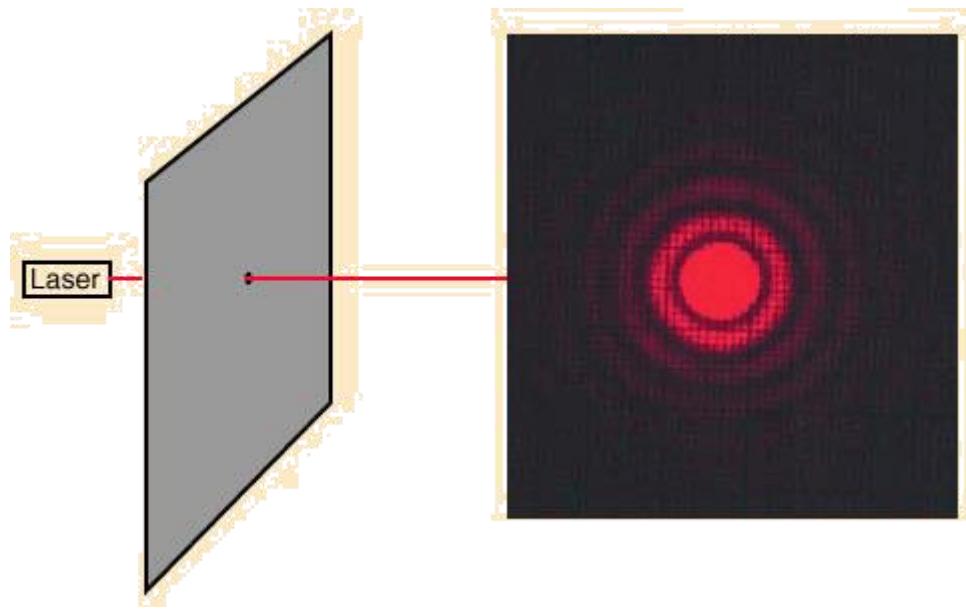


16. Following figure shows the diffraction pattern of a circular aperture. Do you think that the same Huygens's theory of secondary wavelets works here also?
17. Why do Compact Discs appear colored? The pattern changes if viewed at different angles. Why?
18. Can diffraction pattern be unsteady, i.e. time changing? Any example?
19. For diffraction to occur the obstacle should not be too big as compared to the wave being diffracted. Why?
20. The coloration observed on CDs and DVDs is different. Why?
21. Is diffraction applicable to sound waves? If yes, give any day-to-day example

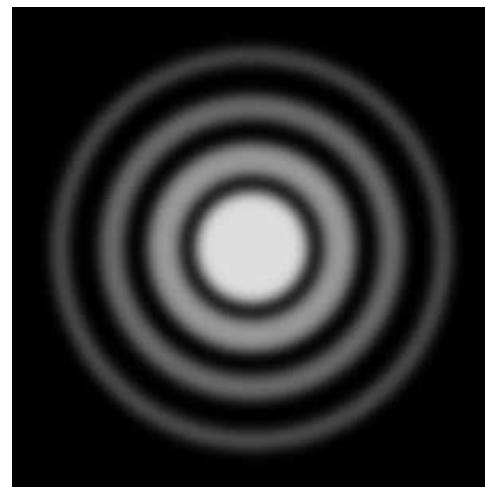
22. Rayleigh's criterion of resolution is tested successfully for several years. Is it an arbitrary or empirical criterion? Can you use the formula for diffraction grating and keep on going mathematically in opposite manner and 'derive' Rayleigh's criterion of resolution?
23. If one wants to verify whether a grating contains really 15000 slits, then what method, based on this chapter itself, he can adopt?
24. Interference requires coherence, but while discussing diffraction, coherence is not emphasized. Why?
25. What should have been the wavelength of light, if a pocket comb having thread spacing was to be used as a diffraction grating and being able to diffract eh first maximum at say 45°
26. Repeat the same calculation for a fenced window having spacing between the rods to be say 30 cm?
27. Both above devices though can't diffract the light, can diffract eh sound waves. Consider the upper and lower wavelength limits of audible spectrum to be 17 m and 0.017 m and calculate the angular position of first maximum if a pocket comb has thread spacing 1 as 1.00 mm and window having spacing between the rods to be 30 cm.
28. If the wavelength of waves in the electromagnetic spectrum are given below then suggest the typical dimensions of the grating element of the corresponding diffraction gratings which can diffract these waves at, say 45° in the first order.



29. In a Giant Meter Wave Telescope (GMRT), near Narayangaon near Pune there are 26 antennas, each antenna having a diameter of 45 m. Why such a big antennas? And why 26 antennas?
30. A grating not only has a limit on the slit density but also on the total number of slits. Why?
31. Can a mirror be converted in to grating? How?
32. Can X- rays be diffracted by the optical gratings? Why? Why not?
33. According to a Nobel laureate Max V. Laue, crystals work as 3-D gratings that can effectively diffract X rays.



34. Fresnel's diffraction is difficult to analyze, while Fraunhofer's one is relatively easy. Why?
35. Do secondary maxima represent the secondary images of the same slit, or do they indicate the well resolved fuzzy edge of the obstacle?
36. Following is the image of a star produced by telescope. If the star is one, why many images?



37. Do you think that the concept of the circular fringes of the Newton's rings or the Michelson's interferometer involved here. Or is this image exclusively due to diffraction?
38. The path difference between the first and the last wave diffracted from the two opposite images of the obstacle have a phase difference as given below. What will be the phase difference between the wave emanating from the upper edge and the middle part of the slit? Should these waves interfere constructively or destructively?

$$\phi = \frac{2\pi}{\lambda} a \sin \theta$$

39. Consider diffraction gratings having 30000 slits per 3 inch, 20000 slits per 2 inch and 10000 slits per 1 inch and calculate their dispersive powers in the first order.
40. Consider diffraction gratings having 20000 slits per inch having total length 2 inch, 20000 slits per inch having total length of 1 inch and 15000 slits having total length of 1.5 inch.
41. Calculate the dispersive powers of the gratings in the above cases.
42. In problems 32, 33 and 34, which grating is best as regards to its resolving power? Which grating is best as regards to its resolving power?
43. A manufacturer wants to produce best grating by accommodating extremely large slits in once inch, but he lands up in a problem, which is not due to technological limitations of the machine. What's the problem then?
44. A grating having maximum dispersive power is to be used. What is it's grating element should be? The average wavelength of light is 5500 \AA .
45. In the following figure two gratings having grating elements of $d = 16933.22 \text{ \AA}$ are used. The first grating is exposed to He-Ne laser having wavelength \AA , which is diffracted at some angle in the first order. This diffracted grating is made to fall normally on the second grating. What is the net angle of diffraction?
46. Three gratings are to be used to give best resolving power. Should these be arranged one on another or one in front of another?
47. Three gratings are to be used to give best dispersive power. Should these be arranged one on another or one in front of another?
48. What do you expect to happen, if the spectrum produced by one grating is allowed to pass through a grating?
49. What do you expect to happen, if the spectrum produced by one grating is allowed to pass through a prism?
50. When a grating is exposed to Mercury, a colored spectrum is produced. When the same grating is used for Helium, Hydrogen, Cadmium or Krypton, the characteristics of spectrum such as the colors, the number of lines or their intensity, everything changes. Why different elements produce different spectra?
51. Somebody claims that, when he went to the market for purchasing the sodium source, he could identify with the help of grating that the source given to him was not sodium. How?
52. In the following Fig a CD is shown, which appears colored due to diffraction effects. The nominal track separation on a CD is 1.6 micrometers, corresponding to about 625 tracks per millimeter. This is in the range of ordinary laboratory diffraction gratings. For red light of wavelength 6000, what should be the angle of diffraction in first order?



REFERENCE BOOKS

9. Fundamentals of Physics Extended, David Halliday, Robert Resnick, Jearl Walker,, John Wiley & Sons
10. The Feynman Lectures on Physics (3 Volume Set), by Richard Phillips Feynman (Author), Robert B. Leighton (Contributor), Matthew Sands (Contributor), The New Millennium Edition, Pearson Education India
 - Excellent websites on this book
 - iii. www.feynmanlectures.caltech.edu/
 - iv. www.feynmanlectures.info/
11. A Textbook of Engineering Physics, M N Avadhanulu & P G Kshirsagar, 10th Edition, S. Chand and Company
12. Fundamentals of Optics, by Francis Jenkins, Harvey White , Tata Mcgraw Hill Publishing Co Ltd
13. Optics, Ajoy K. Ghatak, 5th Edition, McGraw Hill Education,
14. Optics, Eugene Hecht, 4th edition, Addison-Wesley
15. M. Born and E. Wolf, Principles of Optics, Cambridge University Press
16. A Text Book Of Optics, Brijlal, Dr. N. Subrahmanyam, Dr. M. N. Avadhanalu, 25th Edition, S. Chand and Company
 - i.

WORLD WIDE WEB

5. <https://www.photonics.com/>
6. SPIE - the international society for optics and photonics: spie.org/
7. Optical Society of America (OSA): <http://www.osa.org/>
8. Optical Society of India: www.osiindia.org/

CHAPTER 3

Polarization



The first four images are the photographs of LCD screen of a mobile phone taken through a polarizer with different orientations. The difference in the intensities indicates that the light emitted from LCD is polarized. How? The fifth image is a unique application of polarization, called as photoelasticity where distributions of stresses and strains within materials and devices can be visibly observed. Polarization has some other notable applications such as polarizing sun glasses, 3 D movies, Sacirrimetry, LED and LCD monitors etc. What is polarization and how does it lead to these applications?

The answer to this question is in this chapter

Index

3.1 INTRODUCTION

Polarization means restricting the vibrations of light

3.2 TYPES OF POLARIZED LIGHT

Types of restrictions are different

3.3 LAW OF MALUS

Polarization leads to intensity reduction

3.4 POLARIZATION BY REFLECTION

The light reflected at polarizing angle is polarized

3.5 DOUBLE REFRACTION

Where E and O rays come from

3.6 GEOMETRY OF CALCITE CRYSTAL

How to identify optic axis and principle plane

3.7 NICOL PRISM

Eliminating O ray by total internal reflection

3.8 POLAROIDS

Eliminating O ray by selective absorption

3.9 HUYGEN'S THEORY OF DOUBLE REFLECTION

3.10 THEORY OF CIRCULARLY AND ELLIPTICALLY POLARIZED LIGHT

Deriving a general equation of ellipse

3.11 RETARDATION PLATES

Either E or O ray is retarded leading to a path difference

3.12 QUARTER WAVE AND HALF WAVE PLATES

They create a path difference of $\lambda/4$ or $\lambda/2$ between E and O rays

3.13 PRODUCTION OF CIRCULARLY AND ELLIPTICALLY POLARIZED LIGHT

Superimposing E and O rays with path difference of $\lambda/4$

3.14 DETECTION OF POLARIZED LIGHT

The response of a polarized light to a system of polarizer and quarter wave plate is unique to its type

3.15 OPTICAL ACTIVITY

Certain materials can rotate PPL

3.1 INTRODUCTION

Polarization means restricting the vibration of light

Unlike LASER, the ordinary light is emitted by spontaneous deexcitations of atoms. Such deexcitations are random and uncoordinated and therefore the wave trains or the electric and magnetic fields of the such light vibrate in all possible directions perpendicular to the direction of propagation. The number of wavetrains are extremely large and therefore they are isotropically distributed in all directions perpendicular to the direction of propagation. Such light is called as UnPolarized Light (UPL). The vibrations of UPL can be restricted in number of ways leading to Plane Polarized Light (PPL), Circularly Polarized Light (CPL), Elliptically Polarized Light (EPL) and Partially Polarized Light (PRPL). Polarization reveals transverse character of light, as longitudinal vibrations can not be polarized. Unlike interference and diffraction, which is possible for transverse as well as longitudinal waves, polarization is possible only for transverse waves. There are number of devices which work on the principle of polarization. These include Nicol prism, Polaroid, Quarter Wave Plate (QWP), Half Wave Plate (HWP), Polaroscope etc. Based on these devices, there are a few notable applications of polarization such as polarizing sunglasses, Saccimetry, LCD, 3-D movies, Photoelasticity etc. This chapter aims at study of polarization, principles involved in polarizing the light, the polarizing devices and a few applications of polarization. Though light, being an electromagnetic wave consists of vibrations of both electric and magnetic fields, while discussing polarization, we consider only electric vibration as most of the detectors are sensitive to electric field. The polarizing devices that will be discussed in this chapter are based on the optical principles like (i) reflection (ii) transmission through pile of plates (iii) dichroism (selective absorption) (iv) double refraction (v) scattering. The unpolarized and polarized light can not be distinguished by using unaided eye. It requires special devices.

Mechanical analogy for understanding polarization:

Consider a string tied to a rigid support and held in a hand at another end. If the string is oscillated in vertical direction then the wave thus generated is said to be vertically polarized. Such vibrations are represented by an equation

$$y(z, t) = a \cos(kz - wt + \phi_1)$$

If the string is oscillated in the horizontal direction, then the corresponding wave is said to be horizontally polarized. It can be represented by an equation

$$x(z, t) = a \cos(kz - wt + \phi_2)$$

If the hand is rotated in a circle, then the corresponding wave is said to be circularly polarized. It can be represented by an equation

$$x^2 + y^2 = a^2$$

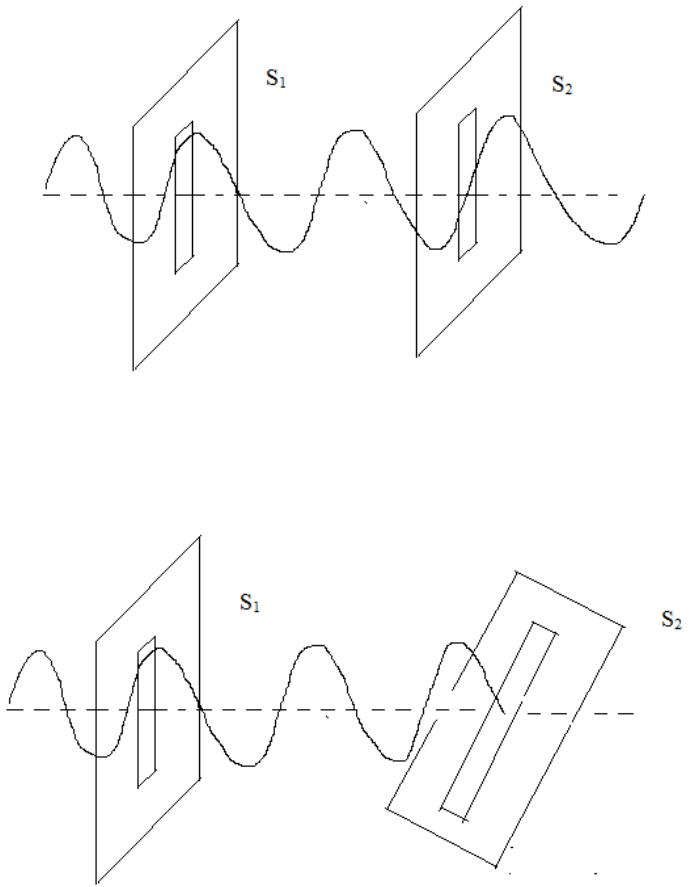


Figure (3.1) : Polarization of waves: Mechanical analogy

If the hand is rotated in an ellipse then the corresponding wave generated on the string is said to be elliptically polarized. It can be represented by an equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

If the string is oscillated in the random directions accessing all the directions perpendicular to the length of the string, then such waves isotropically distributed in all the directions perpendicular to the direction of propagation are said to be unpolarized. In another language, the vibrations which are not restricted in any direction are called as unpolarized vibrations, while the vibrations which are restricted in certain manner are called as polarized vibrations. In case of unpolarized vibrations, the vibrations will be transmitted even if the slit S1 is rotated in all the directions, as for any orientation of the slit, there will be a vibration passing through it. If the slit S1 is not rotated and held fixed in a particular orientation, then it will pass only those vibrations which are parallel to its orientation. Now consider slit S2. When S2 is parallel to S1 it completely passes

the vibrations which are parallel to S1 as well as S2. Now when S2 is rotated with respect to S1, then the intensity of vibrations passing thorough S2 is decreased. When S2 becomes perpendicular to S1, no vibration is transmitted. In another language, when the angle between S1 and S2 is 0° , 180° or 360° , the vibrations are fully transmitted by S2, while if the angle between S1 and S2 is 90° and 270° , then the vibrations through S2 are completely extinguished. Such results will not be obtained if the string is vibrated along its length. In such case the vibrations are said to be longitudinal. The intensity of the longitudinal vibrations passing through S2 will not vary even if S2 is rotated. This experiment clearly indicates that, if the intensity of vibrations passing through S2 varies from maximum to zero and zero to maximum on rotating S2, then the vibrations are transverse in nature and if the intensity does not vary then the vibrations are longitudinal. The vibrations passing through the slit S1, when it is held fixed are said to be polarized as they take place only in a particular direction. The slit S1 is called as polarizer. As the intensity of vibrations passing through S2 varies, S2 can be used to detect whether the vibrations are polarized or unpolarized. S2 can also be used to detect the direction of the polarized vibrations. Thus S2 is called as analyzer.

If the string is passed through a slit and if the string is vibrating vertically, then if the slit is vertical, then the vibrations are completely passed while if the slit is horizontal, then the vibrations are completely blocked. If the string is randomly oscillated in all the directions then then vibrations are called as unpolarized vibrations. In such case for any orientation of the slit, the vibrations will be transmitted and will not be blocked. If the vibrations are longitudinal then the vibrations will be passed for any orientation of the slit.

Understanding polarization: Optical experiment.

Consider an ordinary light instead of a string and a Polaroid (or a tourmaline crystal) instead of a slit. If the Polaroid P_1 is rotated then the intensity of the light will not vary proving that it is unpolarized. However, the case will be different if light is passed through two Polaroids instead of one. If the second Polaroid is rotated across the first one, then intensity of the light will be extinguished twice (when the polarizing directions of two polarizers are at 90° or 270°) and will be maximized thrice (when the polarizing directions of two polarizers are at angles 0° , 180° and 360°). Same observations can be made if the first polarizer is rotated across the second one. The ordinary light (such as that from a sodium lamp or Sun) contains vibrations in all possible directions perpendicular to the direction of propagation. When such light passes through a Polaroid (polarizer), it contains vibrations only in a particular direction decided by polarizing direction (optic axis) of the Polaroid. This experiment conclusively proves that light is a transverse wave. If light were a longitudinal vibration, then the intensity would not vary upon varying the orientation of the second Polaroid.

3.1 TYPES OF POLARIZED LIGHT

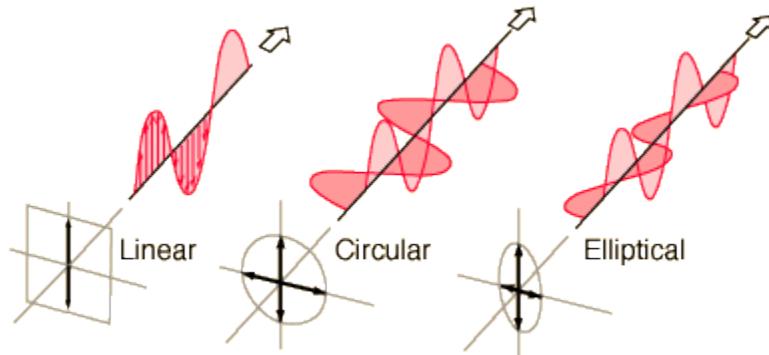
Types of restrictions are different



Figure 3.2 Unpolarized Light (UPL)

Consider Fig. 3.1 which shows unpolarized light. The light vector isotropically vibrates in all the planes perpendicular to the direction of propagation. In unpolarized light, the vibrations are symmetrically distributed in all the directions perpendicular to the direction of propagation. For these vibrations all planes are equally probable. If such light is allowed to pass through a polarizer, then it vibrates only in one direction parallel to its optic axis. As an analogy, optic axis (polarizing direction) can be considered like a slit, which allows only those vibrations that are parallel to the slit. Such light which vibrates only in a particular plane is called Plane Polarized Light (PPL). In polarized light the vibrations are asymmetrically distributed with respect to the direction of propagation.

There are two more kinds of polarized light which are called Circularly Polarized Light (CPL) and Elliptically Polarized Light (EPL). We know that a circular motion is a superposition of two SHMs with a path difference of $\lambda/4$. Thus if two PPLs are superimposed with a path difference of $\lambda/4$ then either CPL or EPL is produced. This is shown in Fig. 3.3. In CPL the superimposing PPLs have equal amplitudes while in EPL, they are unequal. The polarized electric vector in CPL rotates in a circle during its propagation, while in EPL, it traces an ellipse.



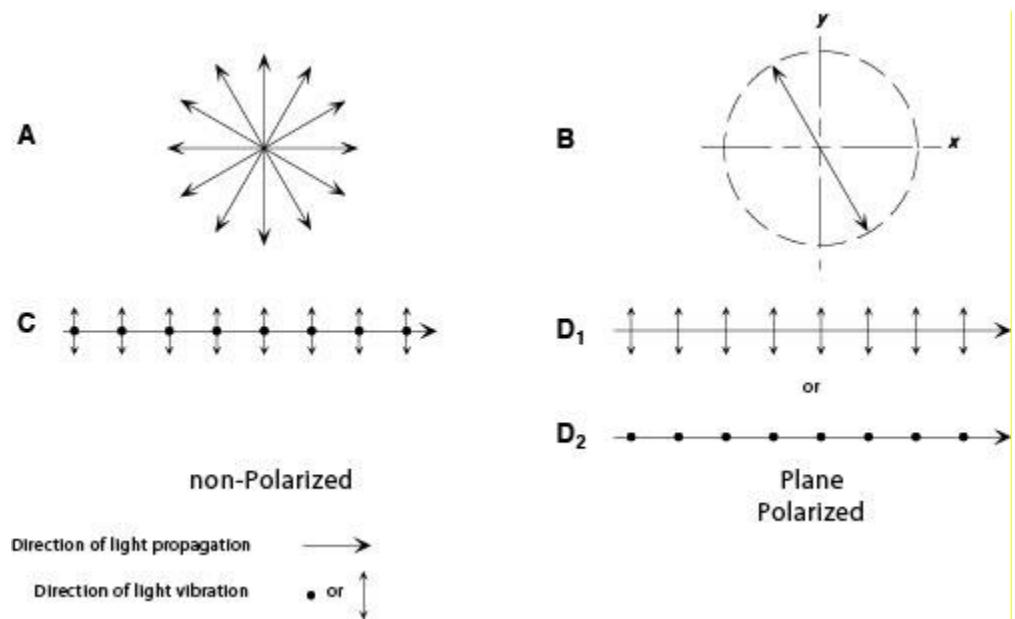


Figure 3.2 Comparison of UnPolarized Light and Plane Polarized Light

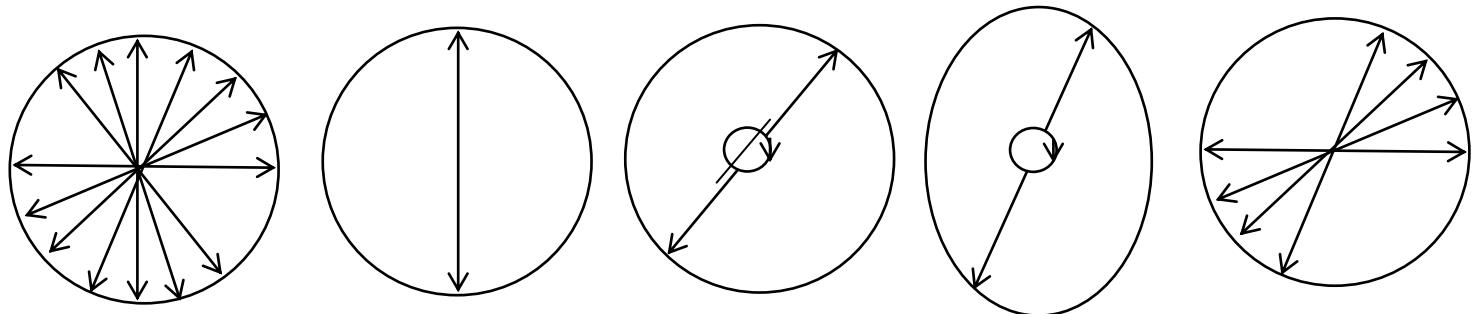


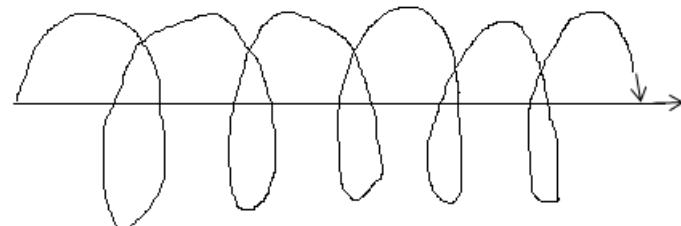
Figure 3.3 a-e UPL

PPL

CPL

EPL

PRPL

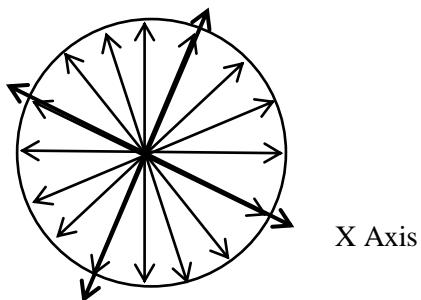


Circularly polarized light

Another representation of circularly polarized light moving towards right side

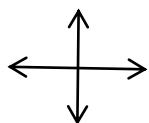
There is one more kind of light, which is neither fully polarized nor fully unpolarized. This is called as partially polarized light.

Intensity reduction during polarization: We know that light wave can be represented by a vector (which is the electric field vector) and the vector can be resolved across the X-Y coordinates. As the unpolarized light has equal vectors distributed across all the directions orthogonal to direction of propagation, after the resolution of all the light vectors along X and Y axis, the sum total of their X and Y components will be equal and will be 50-50%. Now the choice of X and Y axis is a matter of convention. Thus for any orientation of polarizer the Y (or X) axis can be assumed to be along the optic axis of the polarizer, the other axis will be across the optic axis. This indicates that irrespective of the orientation of polarizer, when the light is polarized once, the intensity falls by 50% (in fact the intensity reduction is marginally greater than 50%). As we shall see later, the intensity reduction is greater than 50% if there are two polarizers. In such case the intensity reduction depends upon the angle between the optic axes of the polarizers.

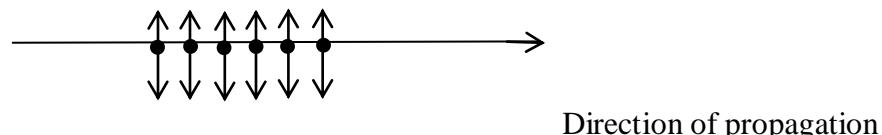


Schematic representations of UPL and PPL

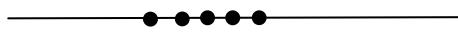
As all the vectors of UPL can be equally resolved along X and Y axes, UPL can be conveniently represented in terms of only X and Y components. If the UPL is coming towards an observer then it can be schematically represented as shown in Fig



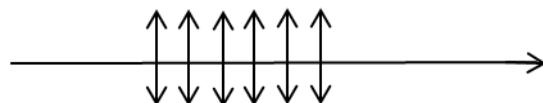
If the light is travelling in horizontal direction, then it can be represented as shown in Fig.



Thus unpolarized light can be represented by a combination of a ‘dot and an arrow’. The vibrations represented by an arrow take place in the plane of incidence, while the vibrations represented by ‘dot’ take place in a plane perpendicular to the plane of incidence. Similarly horizontally and vertically polarized light can be represented by following figures



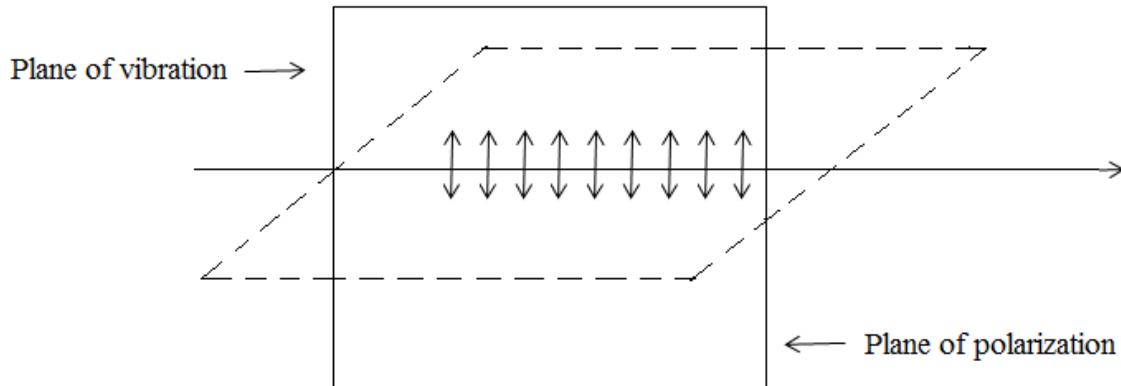
Horizontally polarized light



Vertically polarized light

Plane of vibration: This is a plane in which the vibrations of the plane polarized light take place

Plane of polarization: This is a plane perpendicular to the direction of vibrations of the plane polarized light.



Étienne-Louis Malus (23 July 1775 – 24 February 1812): He was a French Physicist and much of his work is related with polarization. In 1809 he published his discovery of polarization by reflection as he observed sunlight reflected from the windows of the Luxembourg Palace in Paris through an Iceland spar crystal that he rotated. In 1810 he published his discovery of polarization by double refraction by crystals. Malus is probably best remembered for Malus' law, giving the resultant intensity, when a system of polarizers is placed in the path of an incident beam. The term polarization itself is coined by Malus. In 1810 he was awarded Rumford Medal by Royal Society of London. His name is one of the 72 names inscribed on the Eiffel tower.

3.2 LAW OF MALUS (**Derivation is optional**)

Polarization leads to intensity reduction

Consider a monochromatic but unpolarized beam passing through a system of two polarizers.

The angle between their optic axes is θ . Let the amplitude and intensity of original light be E_0 and I_0 . After passing through polarizer 1, the amplitude and intensity are reduced to E_1 and I_1 , where $I_1 = I_0/2$. Now E_1 falls on the polarizer 2 at θ . E_1 can now be resolved along two components, one parallel to and another perpendicular to the optic axis of second polarizer. The parallel component is $E_2 = E_1 \cos \theta$ and the perpendicular one is $E_2' = E_1 \sin \theta$. E_2 will be transmitted and E_2' being perpendicular to optic axis will be extinguished. Thus after the passage of light through the second polarizer, we have

$$E_2 = E_1 \cos \theta$$

Squaring both the sides we have

$$E_2^2 = E_1^2 \cos^2 \theta$$

As Intensity \propto amplitude²,

$$I_2 = I_1 \cos^2 \theta$$

Thus the law of Malus states that the intensity of the light passing through polarizer and analyzer is a cosine square function of the angle between their optic axes. Note that the above equation does not involve I_0 , which is the intensity of original light. At $\theta = 0$, $I_2 = I_1 = I_m$. All rest of the values of I_2 depend upon θ , law of Malus can be expressed in terms of more convenient notation

$$I_\theta = I_m \cos^2 \theta$$

For multiple polarizers, we may write

$$I_3 = I_2 \cos^2 \theta_2$$

$$I_4 = I_3 \cos^2 \theta_3$$

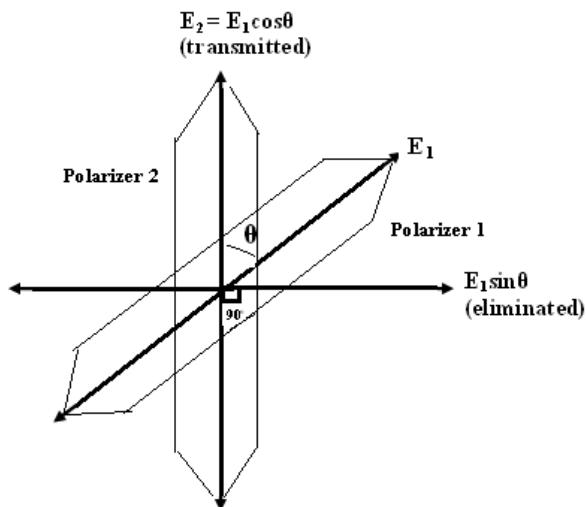


Figure (3.4) Law of Malus

Note that if this law is applied to the first polarizer itself, then the statement will be

$$I_1 = I_o \cos^2 \theta$$

Where I_o is the total intensity of the light incident on the first polarizer, I_1 is the intensity of the light transmitted by the first polarizer. Now as UPL falls on the first polarizer, the electric field vectors which are equally distributed along all the direction subtend all possible angles (θ) with respect to the optic axis of the first polarizer. Therefore an average value of the $\cos^2 \theta$ will have to be considered, which is $\frac{1}{2}$. Thus we get the following expected result

$$I_1 = \frac{I_o}{2}$$

Thus when an unpolarized light is polarized once, its intensity falls by half (if the losses due to absorption are neglected).

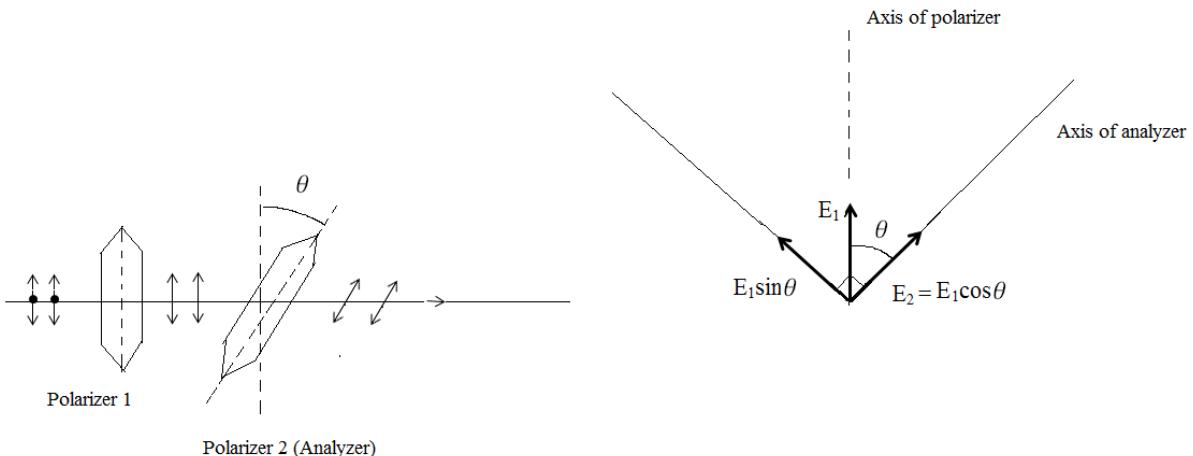


Figure (3.5) Law of Malus: Approach 2

Law of Malus can also be understood by the geometry of Fig . Here the polarizer is vertical while the analyzer is inclined. The angle between them is θ . The component E_1 transmitted through the polarizer can be resolved into two components one is $E_2 = E_1 \cos \theta$, which is along the axis of the analyzer and another is $E_1 \sin \theta$. $E_2 = E_1 \cos \theta$ is transmitted and thus in this case also we can write $I_2 = I_1 \cos^2 \theta$.

Example (3.1)

A polarizer and analyzer are oriented so that the amount of transmitted light is maximum. Through what angle should either be turned so that intensity of transmitted light is reduced to (i) 0.25 , 0.5, 0.75 times the intensity?

Solution:

We have

$$I_\theta = I_m \cos^2 \theta$$

$$\text{Thus } \frac{I_\theta}{I_m} = \cos^2 \theta$$

$$0.25 = \cos^2 \theta$$

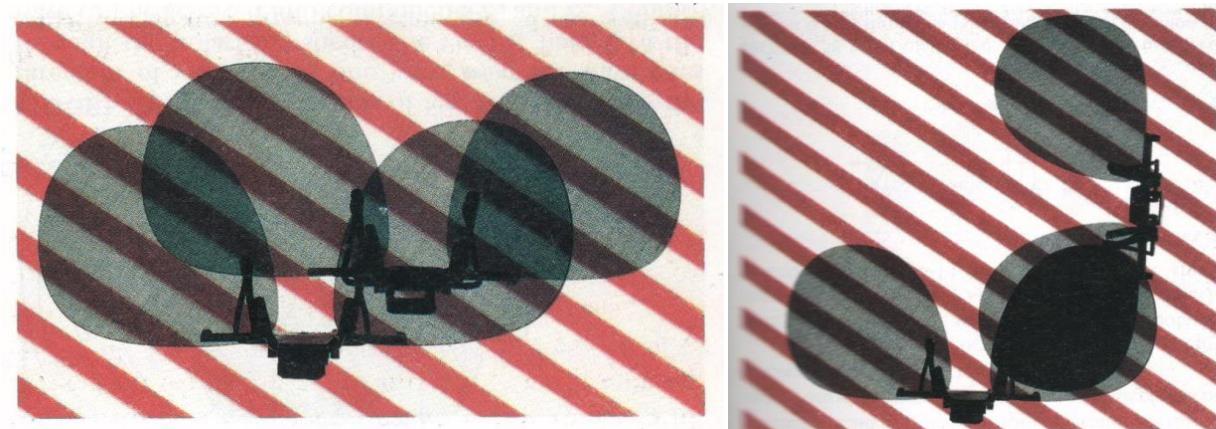
$$\theta = 60^\circ \text{ or } 120^\circ$$

$$0.5 = \cos^2 \theta$$

$$\theta = 45^\circ \text{ or } 135^\circ$$

$$0.75 = \cos^2 \theta$$

$$\theta = 30^\circ \text{ or } 150^\circ$$



Polarizing sunglasses. Law of Malus is clearly demonstrated



Photographs of the LCD screen taken through a polarizer kept at different orientations. This leads to two conclusions, one the LCD emits polarized light and second, it obeys Law of Malus



Another clear demonstration of the law of Malus. Photographs of the sodium source taken through a polarizer rotated at different angles

Example (3.2)

Two polarizing plates have polarizing directions parallel so as to transmit maximum intensity of light. If either of the plate is rotated through 54.74° or 125.26° , by what fraction the intensity through second polarizer will be reduced?

Solution:

We have

$$I_\theta = I_m \cos^2 \theta$$

$$\text{Thus } \frac{I_\theta}{I_m} = \cos^2 \theta$$

$$\frac{I_\theta}{I_m} = \cos^2 54.74^\circ$$

$$\frac{I_\theta}{I_m} = 0.333$$



Sir David Brewster (1781 – 1868). He was a Scottish Physicist. He constructed telescope when he was ten years old. At the age of 19 he was awarded Master's degree by University of Edinburgh. He made many contributions such as polarization by reflection and refraction, discovery of crystals with two optic axes, connection between refractive index and polarizing angle (referred as Brewster's law), invention of the kaleidoscope, invention of the stereoscope etc. In 1815 he became member of Royal Society of London and in 1818 received the Rumford Medal of the Society

3.4 POLARIZATION BY REFLECTION: BREWSTER'S LAW(Optional)



The light reflected at polarizing angle is polarized

Perhaps the simplest method of polarizing the light was discovered by E.L. Malus in 1809. According to this method, if the light is incident on the glass at a specific angle (57°), then the reflected light is plane polarized. Consider Fig where a medium, say a glass is exposed to unpolarized light. The light is incident at arbitrary angle. The reflected light and refracted light are unpolarized. However, the the engle of incidence is increased, then at certain angle called polarizing angle, the reflected light is completely polarized. i.e. it contains only the 'dotted' vibrations, the 'line' vibrations are completely absent. The incident unpolarized light contains 50% 'dotted' vibrations and 50% 'line' vibrations. At polarizing angle, the reflected vibrations contain 15% of the incident 'dotted' vibrations, 85% 'dotted' vibrations and 100 % line vibrations are transmitted. Thus at polarizing angle the reflected light is completely polarized and trasnmitted light is partially polarized.

Consider Fig . A beam of unpolarized light is incident on the first glass plate at 57° . The reflected polarized light is incident on the second glass plate parallel to the first one. As both the glass plates are parallel, the polarized lighth falls on the second glass plate at 57° . In such situation the second glass plate reflects the polarized light. Now, as the second glass plate is rotated with respect to first one, the intensity of the light reflected by the second glass plate starts diminishing. When the second glass plate subtends angles 90° and 270° with the first glass plate, the light reflected by the second glass plate is extinguished complpetely. At the angles 0° , 180° and 360° , the intensity of the light transmitted by the second glass plate is maximum. Conventionally, the first glass plate can be called as polarizer and the second glass plate is called as analyzer. A system of such glass plates thus obeys law of Malus. This system of two paarallel glass plates is called as Biot's polariscope. The polarizing angle varies with the kind of glass plate used.

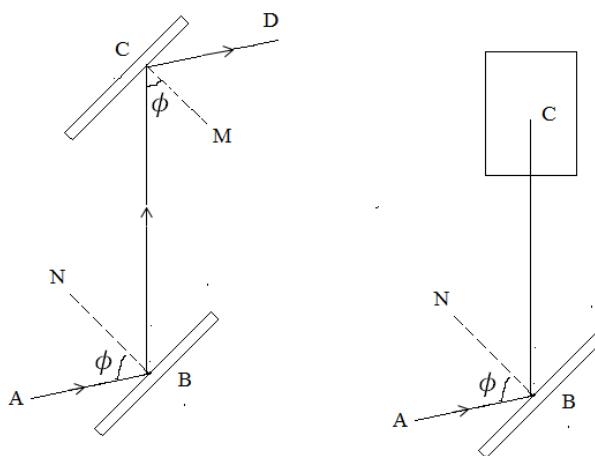


Figure (3.6): Polarization by reflection from glass surfaces



Polarization by reflection: Biot's polariscope

Brewster's law

In 1815, David Brewster discovered that at polarizing angle, when the reflected light is completely polarized, it is completely perpendicular to transmitted light. Thus according to Brewster, at polarizing angle

$$i_p + r_p + 90 = 180$$

Thus

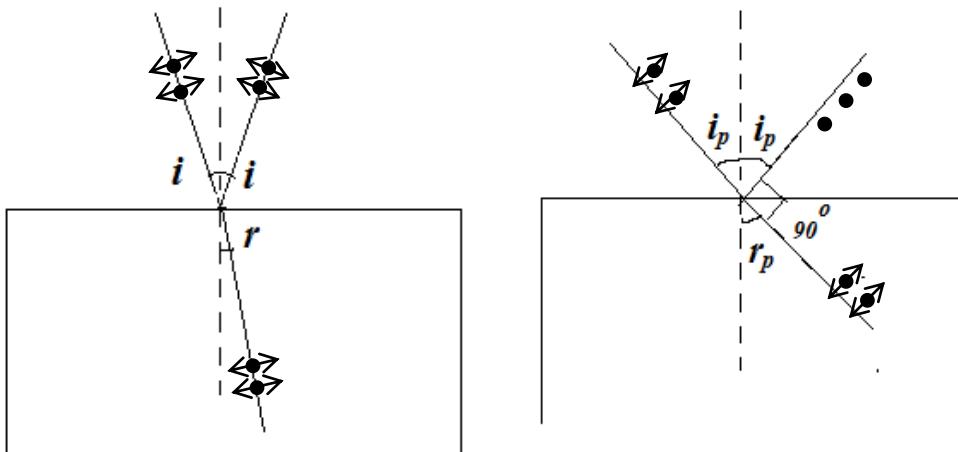
$$i_p + r_p = 90 \quad \dots \text{Brewster's law}$$

According to Snell's law

$$\mu = \frac{\sin i}{\sin r}$$

At Polarizing angle (also called Brewster angle)

$$\mu = \frac{\sin i_p}{\sin r_p}$$



Referring Brewster's law

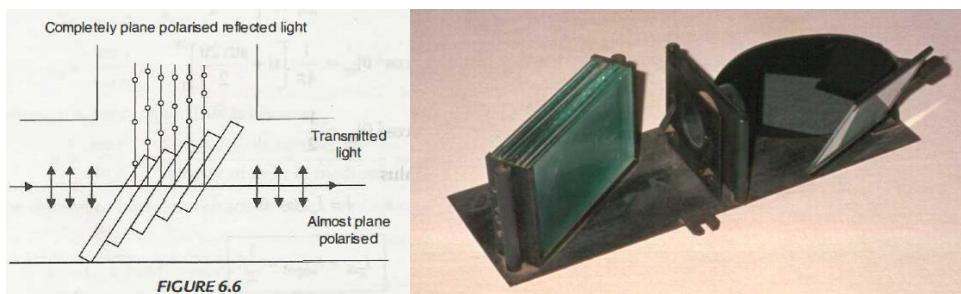
$$\mu = \frac{\sin i_p}{\sin(90 - i_p)}$$

$$\mu = \frac{\sin i_p}{\cos i_p}$$

$$\mu = \tan i_p$$

Above relation signifies that if the polarizing angle of a material is known then its refractive index can be calculated a vice versa. The materials which obey this phenomenon are glass and water.

Polarization by pile of plates: Consider a system of glass plates arranged parallel to each other. We know that, when the light falls on the glass at the polarizing angle, reflected light is completely polarized 15 % of the 'dotted' vibrations are reflected. Thus when the light falls on each successive glass plate at polarizing angle, 15 % of the 'dotted' vibrations are reflected out



Polarization by pile of plates; schematic and the actual device

and the their percentage successively decreases in the transmitted light, the transmitted light contains zero percent ‘dotted’ vibrations and hundred percent ‘line’ vibrations. Thus light can be transmitted by a pile of plates during transmission mode also.

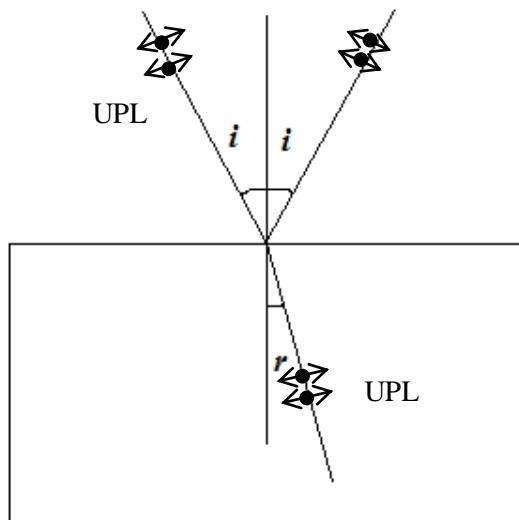
Devices for polarizing and detecting the polarized light: The devices that are used to polarize the light are called as polarizers. There are two examples of polarizers. One is Nicol prism and second is Polaroid. These can also be used as analyzers. Nicol prism and Polaroid are used to produce plane polarized light (PPL). For production of CPL or EPL, it is needed to superimpose two orthogonally vibrating PPLs with a path difference of $\lambda/4$. It may be noted that the vibrations of “E” ray and the “O” ray are in perpendicular planes. The device which produces a path difference of $\lambda/4$ between “E” ray and the “O” ray is called Quarter Wave Plate (QWP). Half Wave Plates (HWP) are also possible. All the devices mentioned in this section are based on double refraction.

3.3 DOUBLE REFRACTION(Compulsory)

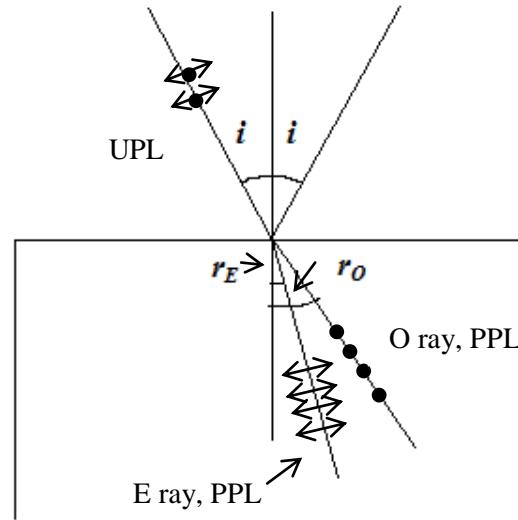


Where E and O rays come from

Refer Fig We know that glass obeys Snell's law. We also know that for glass for a single incident beam there is a single refracted beam. This means that glass is monorefringent. As the angle of incidence varies, the angle of refraction varies according to Snell's law. The refractive index is isotropically same in all



Glass (Monorefrigent)



Calcite (Birefrigent)

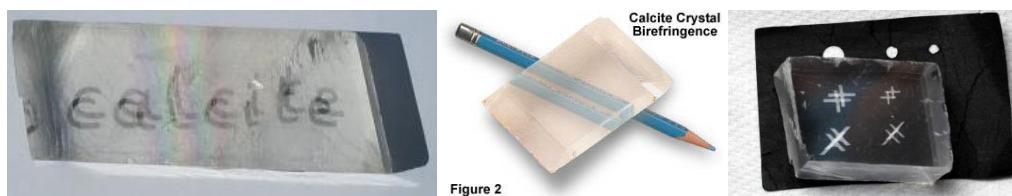
directions. However, in nature, there are some materials which are birefringent. This means that for single incident beam, there are two refracted beams. The examples of this category are calcite, quartz, tourmaline, ice etc. These materials have two refractive indices. (Refer table 3.1). For

two refracted rays, there are two refracted beams. Interestingly, the vibrations of one out of these beams are perpendicular to the plane of paper ('dotted' beam). This ray is called "O" (ordinary) ray. The vibrations of the other beam are in the plane of paper ('line' beam). This ray is called as "E" (extraordinary) ray. For doubly refracting (birefringent) crystals there are two equations of Snell's law as given below.

$$\mu_O = \frac{\sin i}{\sin r_O} \quad \& \quad \mu_E = \frac{\sin i}{\sin r_E}$$

For "O" ray, as angle of incidence is varied, the angle of refraction also varies but by following Snell's law. Thus the refractive index of a birefringent crystal for "O" ray, r_O isotropically remains same in all directions. This is why "O" ray is called as ordinary ray, as it obeys Snell's law. However, for "E" ray, as angle of incidence is varied the angle of refraction r_E varies unisotropically i.e. by not obeying Snell's law. Thus for birefringent crystals, the refractive index of E ray, r_E varies with direction. This is why "E" ray is called extraordinary ray as it does not obey Snell's law.

Devices for polarizing and detecting the polarized light: The devices that are used to polarize the light are called as polarizers. There are two examples of polarizers. One is Nicol prism and Polaroid. These can also be used as analyzers. Nicol prism and Polaroid are used to produce plane polarized light (PPL). For production of CPL or EPL, it is needed to superimpose two orthogonally vibrating PPLs with a path difference of $\lambda/4$. It may be noted that the vibrations of "E" ray and the "O" ray are in perpendicular planes. The device which produces a path difference of $\lambda/4$ between "E" ray and the "O" ray is called Quarter Wave Plate (QWP). Half Wave Plates (HWP) are also possible. All the devices mentioned in this section are based on double refraction.



Demonstration of double refraction. Double image of the letters is seen when viewed through calcite crystal

Crystal	Formula	M_E	μ_O	Birefringence $\Delta\mu = \mu_E - \mu_O$	Type of crystal
Calcite	CaCO_3	1.486	1.658	-0.172	-ve
Ice	H_2O	1.313	1.309	+0.004	+ve
Quartz	SiO_2	1.553	1.544	+0.009	+ve
Siderite	FeO.CO_2	1.635	1.875	-0.240	-ve
Sodium Nitrate	NaNO_3	1.336	1.587	-0.251	-ve

Table 3.1 : Birefringence of various doubly reflecting crystals

Refer table 3.1. The quantity $\Delta\mu = \mu_0 - \mu_E$ is called birefringence and accordingly conventionally, the birefringent materials are classified into -ve and +ve. The materials for which $\Delta\mu$ is -ve i.e. $\mu_0 > \mu_E$ (i.e. $v_0 < v_E$) are called -ve. The examples of this category are Calcite and Sodium nitrate. The crystals for which $\Delta\mu$ is +ve i.e. $\mu_0 < \mu_E$ (i.e. $v_0 > v_E$) are called +ve. The examples of this category are Quartz and Ice.

3.4 GEOMETRY OF CALCITE CRYSTAL (Optional)



How to identify optic axis and principle plane

Though there are a few doubly refracting crystals, Calcite is the most preferred one, as it has certain advantages. First is that it is a naturally occurring crystal, it is relatively cheap. It is colorless, transparent and easy to cut. A polarizer called Nicol prism is made using Calcite. Therefore it is necessary to understand the geometry of Calcite. Calcite is an example of anisotropic crystal in the sense that its physical properties vary with the direction. Its geometry is hexagonal. It's uniaxial crystal that means it has only one optic axis, the axis of symmetry as regards to crystal form. As shown in Fig. point B and B' are two blunt corners as at these points meet at obtuse angles. Optic axis of calcite is the line joining B and B'. At B and B' the faces meet the optic axis at equal angles. Any doubly refracting crystal exhibits no birefringence along the optic axis. It is to be noted that optic axis is a direction and not a concrete line. Thus any line passing through the crystal and parallel to the line BB' behaves as an optic axis of the crystal. Principle plane of Calcite is a plane, containing line BB' (i.e. the optic axis) and perpendicular to any cleavage face. A principle section always cuts the surfaces of the calcite crystal in a parallelogram with angles 71° and 109° as shown in the Fig.

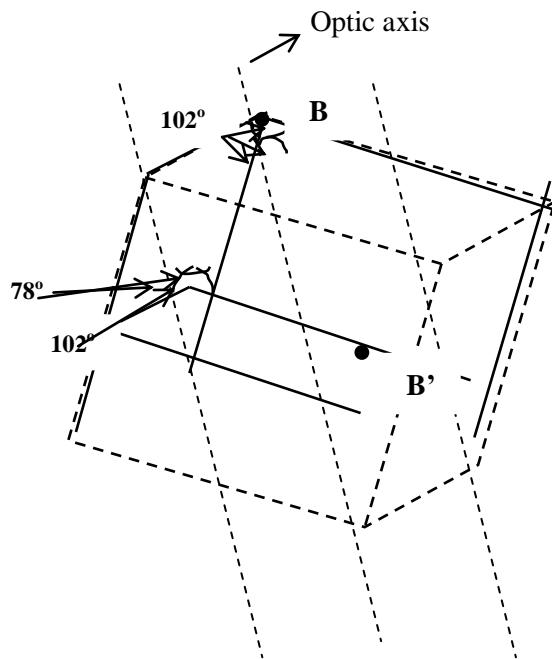
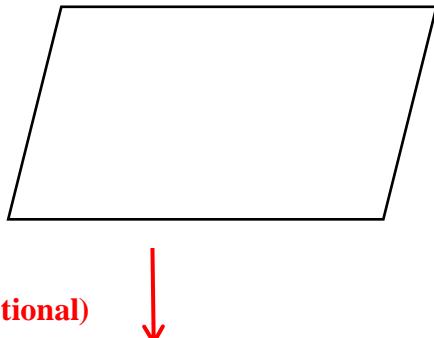


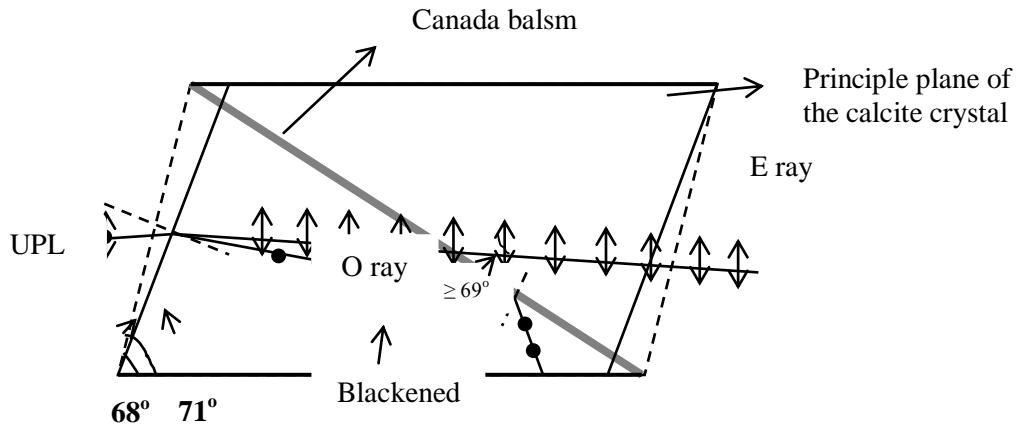
Figure (3.6) Geometry of Calcite crystal



3.5 NICOL PRISM (Optional)

Eliminating O ray by total internal reflection

Nicol prism was constructed by a Scottish Physicist William Nicol in 1828. For its construction, a rhombohedral calcite slab with length 3 times its breadth is taken. Refer Fig. It shows Nicol prism as regards to its principle plane is shown. One of the angle 71° is reduced to 68° by cutting it appropriately. Then it is cut along shorter diagonal a divided into two equal halves. These halves are then pasted by an optical paste called *Canadabalsm*. Canada balsm is a transparent paste with refractive index 1.55 which is same for both E and O rays. Calcite is a -ve crystal with its refractive index 1.486 for E ray and 1.658 for O ray. Thus Canada balsm is optically denser for E ray and rarer for O ray. Thus if O ray falls on Canada balsm at an angle greater than equal to critical angle then it will be totally internally reflected and will not be transmitted further.



The critical angle for O ray can be calculated by eqn.

$$\theta_c = \sin^{-1} \frac{1.55}{1.658}$$

$$= 69.20^\circ$$

When a monochromatic beam of UPL is incident on Nicol prism, it is doubly refracted in to E and O rays. As shown in the Fig. both these rays fall on Canada balsm. Canada balsm being denser for E ray, total internal reflection is not possible and E ray is transmitted. O ray falls on Canada balsm at angle greater than or equal to 69° and is totally internally reflected, its

transmission is blocked. After total internal reflection, O ray falls on the surface which is blackened and thus absorbed.

Can the total internal reflection of O ray take place arbitrarily at all the conditions? No! For total internal reflection to be made possible, the length is made 3 times the breadth, the angle of principle plane is decreased from 71° to 68° . Further the angle of incidence has to be less than or equal to 14° .

Parallel and crossed Nicols: Nicol prism can be used as polarizer as well as analyzer. When two Nicol prisms are parallel, their principle planes are parallel, then the E ray passes through second Nicol prism also. However, when the Nicol prisms are crossed, the E ray emanating from first Nicol prism behaves as O ray in second Nicol as the principle planes are rotated. This ray falls on the Canada balsm of second Nicol at greater than 69° and is totally internally reflected. Thus two parallel Nicols transmit the light but light is extinguished when Nicols are crossed.

3.6 POLAROIDS(Optional)



Eliminating O ray by selective absorption

Tourmaline is a doubly refracting crystal. However, it has a natural property of selectively absorbing O ray. This property of selective absorption of O ray is called dichroism. Thus only E

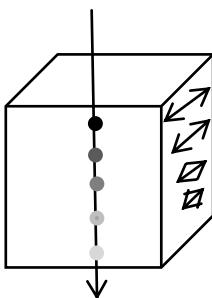
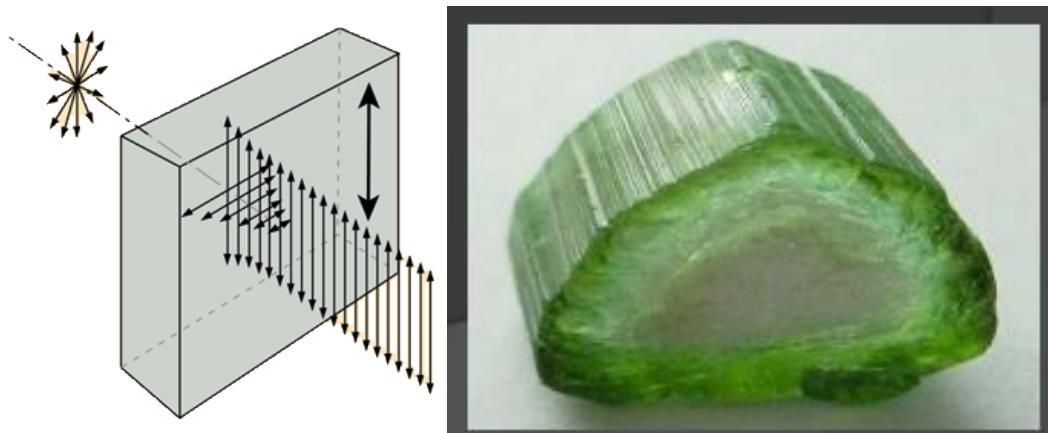


Figure (3.8) Selective absorption of O ray



Selective absorption in tourmaline: Schematic and actual crystal

ray is transmitted and light is polarized. Refer Fig. However, tourmaline is somewhat colored (greenish) and is therefore not preferred in polarizer.



Edwin Herbert Land (1909 –1991): He was an American Physicist, He studied at Harvard University. He is best known for Polaroid corporation. Among the other things he invented are inexpensive filters for polarizing light known as Polaroid film, a practical system for in-camera instant photography and his retinex theory for color vision. His Polaroid instant camera made it possible for a picture to be taken and developed in 60 seconds or less. His Polaroid films were used for sunglasses and photographic filters, glasses in full-color stereoscopic (3-D) movies, to control brightness of light through a window, a necessary component of all LCDs, and many more. What is discussed in forthcoming paragraph is Polaroids, which was invented by E.H. Land when he was a 19 year old student at Harward University

W. B. Herapath discovered in 1852 a synthetic dichroic crystal quinine iodosulfate (now known as herapathite). Herapathite has a property of selective absorption of O ray. However, till 1932, a method to synthesize large crystals of herapathite was not known. E. H. Land invented in 1928 of aligning crystals to obtain an efficient and economic polarizing material. One of the method to obtain such crystals is to crush herapathite in to ultrafine powder. The microscopic crystals in such powder have random distribution of optic axes. The powder is then dissolved into an optical paste of nitrocellulose. The paste is then pressed through a fine slit so that the optical axes of most of the microscopic crystals in the paste are aligned. The thin film aligned between glass plates becomes a polarizer.

Land and Rogers further developed a method to synthesize inexpensive sheets of polarizers based on a plastic called Polyvinyl Alcohol (PVA). There are two kinds of such Polaroids; K-Polaroid and H-Polaroid. In K-Polaroid, PVA is stretched to many times its original length. In this process its molecules are aligned. The stretched PVA is then impregnated with iodine. The stretched axis along which iodine is impregnated becomes conducting. When a UPL is passed through such PVA, the vibrations parallel the iodinated axis are transmitted while the vibrations across the iodinated axis are eliminated. The emerging light thus plane polarized. In H-Polaroid the PVA is heated in presence of a dehydrating agent such as HCL. This PVA becomes somewhat colored but acquires a strong dicroism and behaves as a polarizer.



Christiaan Huygens, FRS (1629 – 1695). He was a prominent Dutch Physicist. His work included telescopic studies elucidating the nature of rings of Saturn using his 50 power refracting telescope, discovery of its moon Titan, observation of Orion nebula, the inventions of pendulum and spring clock, studies of Optics, centripetal and centrifugal force etc. He also played a role in the development of modern Calculus. He became the member of Royal society in 1663. Huygens is especially remembered for his wave theory of light, now known as Fresnel-Huygens principle which he published in *treatise on light* in 1690, one of his few publications. Note that his wave theory of light was published much before Thomas Young who published it in 1801. He is credited as the inventor of first Magic Lantern. He also designed accurate clocks based on pendulums and springs and also published theory of coupled oscillations. He was the first to derive the formula for period of pendulum. He also designed internal combustion engine. He had also developed a technique for accurate measurement of astronomical distances. He was a believer of extraterrestrial life. Huygens experimented with double refraction in Calcite and explained it with wave theory and polarized light. What is presented in next few paragraphs is his explanation of double refraction using theory of wavefronts

3.7 HUYGEN'S THEORY OF DOUBLE REFLECTION(**Compulsory**)

Visualizing E and O ray in the form of wavefronts



We know that according to Christiaan Huygens, light energy propagates in the form of wavefronts. This must be true for polarized light also. We know that there are two kinds of PPLs namely E ray and O ray. We also know that their optical properties are different. It will be of interest to visualize E and O ray in terms of wavefronts. It will be seen later on that such discussion will be quite useful for understanding Quarter Wave Plates and Half Wave Plates. Consider Fig. in which a monochromatic source of light is assumed to be inside the crystal. As light begins to propagate, there will be double refraction. Thus E and O rays will come into existence. We know that O ray obeys Snell's law, and the refractive index of a doubly refracting crystal for O ray is same in all directions.

Thus the velocity of O ray will be same in all the directions ($\mu_o = c/v_o$). Thus the wavefront of O ray is spherical. The section of this wavefront in the principle plane appears circular. The circular nature of wavefront of O ray will be same in both negative and positive crystal. The case of E ray is different. E ray does not obey Snell's law and refractive index of a doubly refracting crystal varies with the directions. All doubly refracting crystals obey one property. Their birefringence is zero along the optic axis (axis of symmetry). Thus the refractive index of the crystal for E and O ray are same along Optic axis. Thus E and O ray propagate with same speed along optic axis. Thus their wavefronts touch along the optic axis. This is true for both negative and positive crystals. The birefringence increases as one move away from the optic axis. It becomes maximum in a direction perpendicular to optic axis. Then it gradually decreases

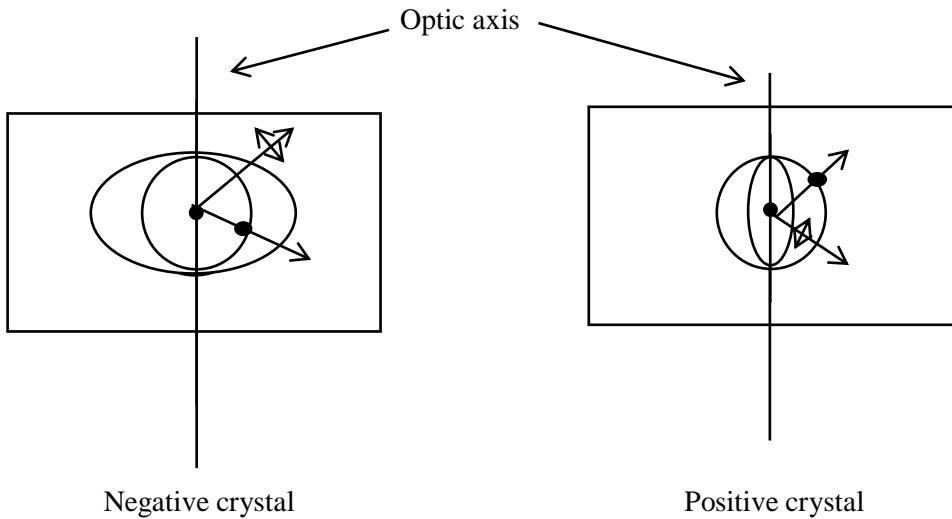


Figure (3.8) Propagation of E and O rays in terms of wavefronts. The monochromatic source of light is inside the crystal

and zero as one approaches towards optic axis. Thus the difference in refractive indices and the velocity between E and O ray gradually increases while approaching towards optic axis, becomes maximum in a direction perpendicular to optic axis and once again gradually decreases while approaching towards optic axis. This is clearly depicted in the Fig which shows maximum path difference between E and O rays along the direction perpendicular to optic axis. The wavefront of O ray as observed in the principle plane is circular while for E ray it is elliptical because the specific variation of refractive index with the direction. The wavefronts of O and E ray touch along the optic axis. For the negative crystal like calcite, where $v_o < v_e$, the circle lies inside the ellipse while for the positive crystal where $v_o > v_e$, the circle lies outside ellipse.

Placing source of light inside the crystal is purely an imagination. Figures and depict the pictures of wavefronts when source of light is outside the crystal. The figures depict three facts, one is that in every case the wavefronts of E and O ray touch along the optic axis, second is that they show maximum path difference in a direction perpendicular to optic axis and third is that for -ve crystal circle is inside the ellipse and for positive crystal circle is outside the ellipse.

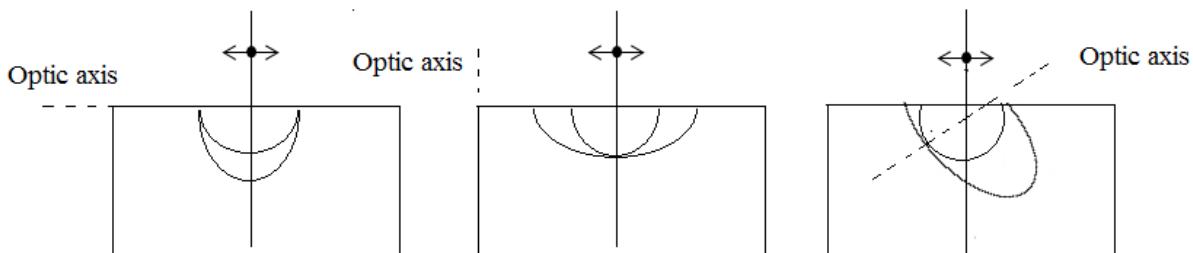


Figure (3.9) The wavefronts of E and O rays for different orientations of optic axis for negative crystal

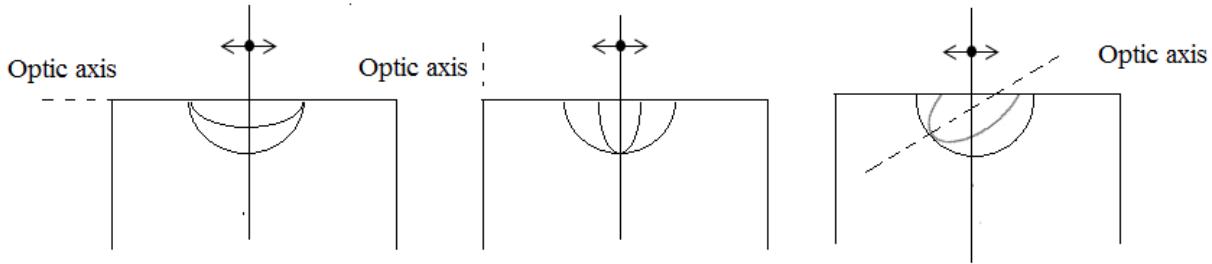


Figure (3.10) The wavefronts of E and O rays for different orientations of optic axis for positive crystal

3.8 THEORY OF CIRCULARLY AND ELLIPTICALLY POLARIZED LIGHT **(Optional)**

Deriving a general equation of ellipse

Consider two sine vibrations represented by

$$x = a \sin \omega t, \quad y = b \sin(\omega t + \phi)$$

Note that x and y can represent E and O ray as they vibrate orthogonally. Our aim is to see what will happen if E and O ray are made to vibrate orthogonally with an arbitrary phase difference ϕ . Thus we need to solve equations for x and y simultaneously.

$$\frac{x}{a} = \sin \omega t, \quad \frac{y}{b} = \sin \omega t \cos \phi + \sin \phi \cos \omega t$$

Substituting $\sin \omega t = \frac{x}{a}$ in $\frac{y}{b}$, we get

$$\frac{y}{b} = \frac{x}{a} \cos \phi + \sin \phi \sqrt{1 - \frac{x^2}{a^2}}$$

$$\left(\frac{y}{b} - \frac{x}{a} \cos \phi \right)^2 = \left(\sin \phi \sqrt{1 - \frac{x^2}{a^2}} \right)^2$$

$$\frac{y^2}{b^2} - 2 \frac{xy}{ab} \cos \phi + \frac{x^2}{a^2} \cos^2 \phi = \sin^2 \phi - \sin^2 \phi \frac{x^2}{a^2}$$

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - 2 \frac{xy}{ab} \cos \phi = \sin^2 \phi \quad \dots (3.3)$$

Eqn (3.3) represents a general equation of ellipse. Thus if two sine vibrations vibrating

orthogonally with a phase difference are made to superimpose with a phase difference of ϕ , it results into elliptical vibration. Let us now discuss various special cases of eqn. (1), with different values of ϕ

ϕ	Resulting eqn	Resulting vibration	Remark
$0^\circ, 360^\circ$	$y = \frac{b}{a}x$	Straight line	If two PPL is superimposed with zero phase difference then resulting vibration is PPL
180°	$y = -\frac{b}{a}x$	Straight line	If two PPL is superimposed with zero phase difference then resulting vibration is PPL
$90^\circ, 270^\circ$ $a \neq b$	$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	Ellipse	If two PPLs with unequal amplitudes are superimposed with a phase difference of $90^\circ, 270^\circ$, the resulting vibration is EPL
$90^\circ, 270^\circ$ $a = b$	$x^2 + y^2 = a^2$	Circle	If two PPLs with equal amplitudes are superimposed with a phase difference of $90^\circ, 270^\circ$, the resulting vibration is CPL

Table (3.2): Special cases of ellipse

The above table depicts methods of producing various kinds of polarized lights. What is only needed is to superimpose x and y (i.e. E and O) vibrations with desired phase differences. All this can be achieved with the help of a device called as retardation plate. There are two kinds of retardation plates, one is Quarter Wave Plate (QWP) and another is Half Wave Plate (HWP)

QUARTER WAVE AND HALF WAVE PLATES(**Compulsory**)

They create a path difference of $\lambda/4$ or $\lambda/2$ between E and O rays

We know that depending upon a negative or positive crystal either O ray or E ray is retarded. This is the basic principle behind retardation plate. While designing the retardation plate, the direction of propagation is kept perpendicular to optic axis, so that maximum path difference is created. Calcite is preferred due to its properties discussed earlier. Calcite has strong birefringence. Retardation plates are of two kinds, one is called as Quarter Wave Plate (QWP) and another as Half Wave Plate (HWP). QWP is designed to create a path difference of $\lambda/4$ between E ray and O ray, while HWP creates a path difference of $\lambda/2$ between these rays.

As shown in fig. consider a calcite slab of thickness "t" desired to create a path difference of $\lambda/4$ between E and O ray. For taking the benefit of maximum birefringence, the direction of

propagation of incident light is kept perpendicular to optic axis. Consider a monochromatic ray

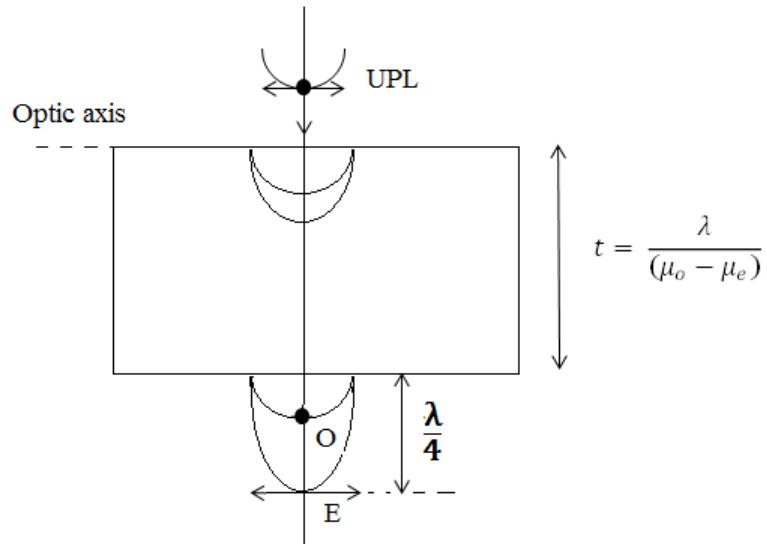


Figure (3.10) Quarter Wave Plate (QWP) made of calcite crystal

of UPL incident on the calcite slab. As the ray enters in the slab, the double refraction will begin. For negative crystal the O ray will start lagging behind E ray. Though the difference between the velocities is constant, the path difference between E and O ray will keep on increasing. The slab should be cut in such a way that when E and O ray emerge out of the slab, O ray will exactly lag behind E ray by $\lambda/4$. The formula for required thickness of the slab can be derived by considering the optical paths of E and O ray through the crystal.

We know that

$$(\text{Optical path}) = \mu (\text{Geometrical path})$$

Thus the optical paths of E and O ray through the slab of thickness t will be $\mu_e t$ and $\mu_o t$ respectively. As the crystal is negative $\mu_o t > \mu_e t$. Thus the path difference will be

$$P.D. = \mu_0 t - \mu_e t$$

For QWP, the desired path difference is $\lambda/4$, thus,

$$\frac{\lambda}{4} = \mu_0 t - \mu_e t$$

$$t = \frac{\lambda}{4(\mu_o - \mu_e)} \quad \dots(3.4)$$

Eq. () gives the desired thickness of QWP of negative crystal. The WQP of positive crystal and HWP of positive and negative crystal can be discussed on similar lines. Table () gives the formulae for QWP and HWP of negative and positive crystals

Device	Negative crystal	Positive crystal
QWP	$t = \frac{\lambda}{4(\mu_o - \mu_e)}$	$t = \frac{\lambda}{4(\mu_e - \mu_o)}$
HWP	$t = \frac{\lambda}{2(\mu_o - \mu_e)}$	$t = \frac{\lambda}{2(\mu_e - \mu_o)}$

Table (3.3): QWP and HWP

Example (3.3)

Calculate the thickness of Quarter Wave Plate of Calcite for sodium light having wavelength 5890 A°. ($\mu_o = 1.658$ and $\mu_e = 1.486$)

Solution

$$t = \frac{\lambda}{4(\mu_o - \mu_e)}$$

$$t = \frac{5890}{4(1.658 - 1.486)}$$

$$t = 8561 \text{ A}^\circ$$

$$t = 0.8561 \mu\text{m}$$

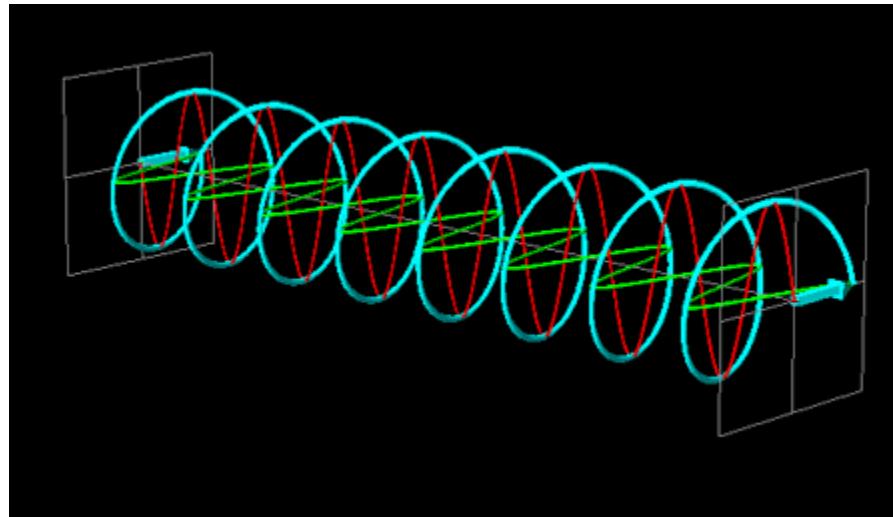
This value of the thickness indicates that such slabs are to be extremely thin and rather should be referred as films.

3.10 PRODUCTION OF CIRCULARLY AND ELLIPTICALLY POLARIZED LIGHT(Optional)

↓ *Superimposing E and O rays with path difference of $\lambda/4$*

We know that circular motion in a combination of two perpendicular SHMs. In section ...we have also seen that if two perpendicular vibrations are made to superimpose with phase difference of 90° (path difference of $\lambda/4$), then the resulting motion is circular. Almost every such condition required to produce CPL is achieved when a QWP is exposed to PPL. If such PPL during its incidence makes an angle of 45°, then its components E and O will have same amplitude. After passing through QWP, E and O rays will vibrate together with equal amplitudes

and path difference of $\lambda/4$ (phase difference of 90°), then resulting light will vibrate circularly and CPL will be produced. If the incident PPL makes an angle other than 0° , 90° or 45° , then the components E and O will have unequal amplitudes and the resulting vibration will be EPL. This is explained in Fig and



3.9 DETECTION OF POLARIZED LIGHT

The response of a polarized light to a system of polarizer and quarter wave plate is unique to it's type

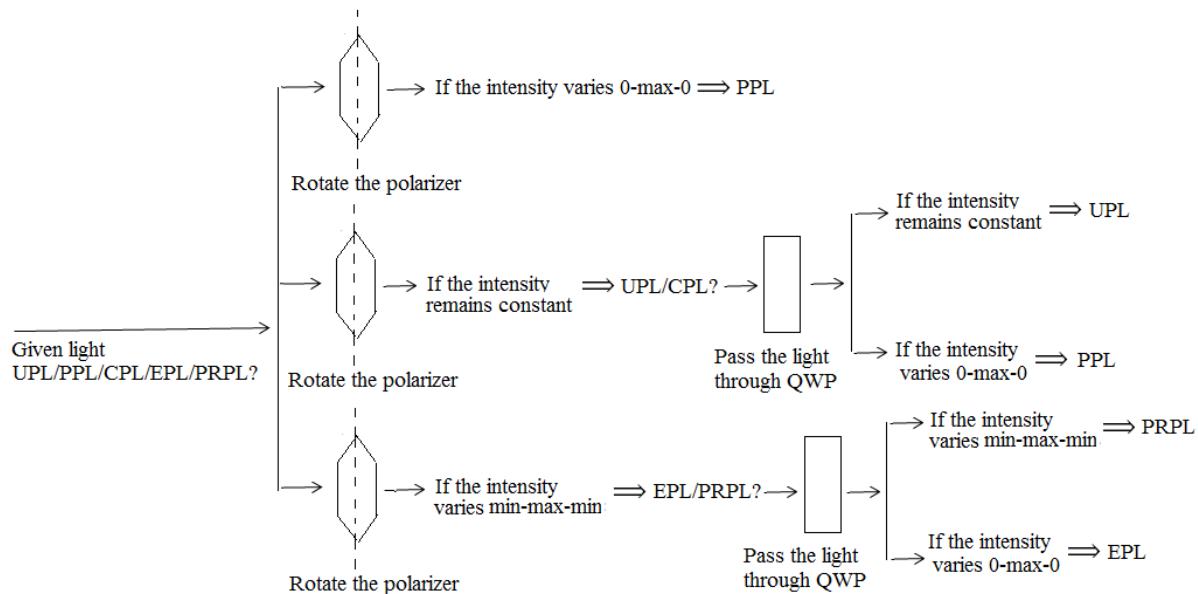


Figure (3.11) Analysis/Detection of kind of polarized light

Consider a beam of light which either UPL/PPL/CPL/EPL or PRPL. A system of polarizer and QWP can be used to detect it's kind. The first test is off course to rotate the polarizer across the beam of light. Three different responses are possible. Either the intensity of light may vary from 0-max-0

Applications of Polarization

- a. Polarizing sunglasses
- b. Visors of automobiles
- c. 3 D movies
- d. Liquid Crystalline Displays
- e. Photoelasticity
- f. Optical Activity

3.10 OPTICAL ACTIVITY(Optional)



Certain materials can rotate PPL

A few materials in the nature have an ability to rotate the plane of vibration when PPL passes through them. This property is called optical activity. The materials which exhibit this property are quartz, cinnabar. A few liquids such as turpentine, tartaric acid, nicotine, aqueous solutions of sugar are the examples of liquids which exhibit optical activity.

There are two kinds of optically active materials. The materials which rotate the plane of vibration in the right hand side or in clockwise direction are called dextrorotatory or d-rotatory. The materials which rotate the plane of vibration in left hand side or in anticlockwise direction are called levorotatory or l-rotatory.

Figure

Crystalline substances exhibit optical activity to largest extent when PPL passes along the optic axis.

The amount of rotation depends upon wavelength of light (rotatory dispersion). The angle of rotation is inversely proportional to the square of the wavelength of light. For a monochromatic beam of light, the angle of rotation θ depends upon the length through which PPL passes through the material, it also depends upon the concentration of the material

$$\theta = \alpha l \quad \dots(1)$$

α is called rotational constant.

$$\theta = s c l \dots(2)$$

The quantity s is called the specific rotation. If a light is passed through parallel polarizers, maximum light will be transmitted. But if an optically active material is held between then polarizers, the plane of vibration of the PPL will be rotated. Now the system of two polarizers will not pass the maximum light. In such case the analyzer will have to be rotated by the same angle θ by which the PPL is rotated. Thus θ can be measured. If the length the optically active

material and its specific rotation are known then the concentration of optically active material can be calculated. This method is widely used in sugar industry for measuring the concentration of sugar in its solution and this technique is called saccarimetry.

The specific rotation s is defined as the rotation when the length of optically active material (or the length of the glass cell containing optically active solution) is 1 decimeter and the concentration of the solution in 1gm/cc.

A satisfactory explanation of the optical activity is given by Fresnel. A PPL can be thought to be a combination of two CPLs rotation in opposite direction. When such PPL passes through a medium the PPL is split into two CPL with half amplitude. In ordinary material exhibiting no optical activity the electric field vectors of both CPLS rotate in opposite directions with equal velocity so that the orientation of the resultant PPL remains unchanged. However, in case of optically active material both these CPLs rotate with different angular velocity. Depending upon which angular velocity is larger the resultant PPL either rotates in clockwise or anticlockwise direction.

TUTORIAL ON POLARIZATION

1. A polarizer and analyzer are oriented so that the amount of transmitted light is maximum. Through what angle should either be turned so that the intensity of light is reduced to (i) 0.75 and (ii) 0.25 times the maximum intensity?
2. Polarizer and analyzer are set with polarizing direction parallel so that the intensity of light is maximum. If either polarizer are rotated by 45° (or 135°) or 60° (120°) then what is the fraction of the intensity reduced.
3. Two polarizers are crossed so that they transmit zero intensity. If the third polarizer is inserted at 45° w.r.t. both, what is the fraction intensity transmitted by both.
4. For ordinary glass plate $\mu=1.54$. Find the angle of polarization and corresponding angle of refraction.
5. It is observed that when the light is incident on the surface of benzene at 56.31° , the reflected light is linearly polarized. What is the refractive index of benzene?
6. At a certain temperature the critical angle of incidence of water for total internal reflection is 48° for a certain wavelength. What is the polarizing angle and angle of refraction for light incident on water such that given maximum polarization of reflected light?
7. At what angle of incidence should a beam of sodium light be directed upon the surface of diamond to produce most complete polarization? The critical angle for diamond is 24.5° .
8. Calculate the thickness of (i) Quarter wave plate and (ii) Half wave plate for quartz, given $\mu_e = 1.553$, $\mu_o = 1.544$, $\lambda = 5000 \text{ Å}$.
9. Calculate the thickness of doubly refracting crystal required to introduce a phase difference of π radians, given that $\lambda = 6000 \text{ Å}$, $\mu_e = 1.553$, $\mu_o = 1.544$, $\lambda = 5000 \text{ Å}$. Also calculate the birefringence of the crystal
10. A 20 cm long tube containing 48 cc of sugar solution rotates the plane of polarization by 11° . If the specific rotation of sugar solution is 66° , calculate the mass of the sugar in the solution
11. 80 gm of impure sugar is dissolved in a liter of water. The solution gives optical rotation of 9.9° when placed in a tube of length 20 cm. If the specific rotation of sugar solution is

$66^{\circ} \text{ dm}^{-1} (\text{gm/cc})^{-1}$. Find the percentage of purity of the sample.

12. A sugar solution in a tube of length 20 cm produces optical rotation of 13° . The solution is then diluted to $1/3$ of its previous concentration. Find optical rotation produced by 30 cm long tube containing the diluted solution.

13. A certain length of 5% solution causes the optical rotation of 20° . How much length of 10 % solution of the same substance will cause 35° rotation.

MULTIPLE CHOICE QUESTIONS

- c Vibrations parallel to the direction of propagation d None of a, b and c
- Q 8** In the phenomenon of polarization by reflection, if the light is incident on a glass plate at the polarizing angle, then the refracted light contains,
- a Vibrations perpendicular to the plane of incidence b Vibrations in the plane of incidence
c Mixture of vibrations perpendicular to plane of incidence and in the plane of incidence d None of a, b and c
- Q 9** Which of the following substance cannot be used for polarization by reflection
- a Benzene b Water
c Glass d None of a, b and c
- Q 10** The refractive indices of crown and flint glasses are 1.485 and 1.523 respectively. Their polarizing angles are
- a 56.04° and 56.71° respectively b 56.71° and 56.04° respectively
c 55° and 60° respectively d None of a, b and c
- Q11** Which of the following material is not doubly refracting?
- a Calcite b Ice
c Tourmaline d None of a, b and c
- Q 12** Birefringence is defined as
- a $\Delta\mu = \mu_o - \mu_e$ b $\Delta\mu = \mu_e - \mu_o$
b $\Delta\mu = \mu_e + \mu_o$ d $\Delta\mu = \frac{\mu_e}{\mu_o}$
- Q13** Calcite is a
- a Rhombohedral crystal b Cubical crystal
c Hexagonal crystal d Triclinic crystal
- Q 14** The formula for the thickness of Quarter Wave Plate made up of positive crystal is
- a $t = \frac{\lambda}{4(\mu_o - \mu_e)}$ b $t = \frac{\lambda}{4(\mu_e - \mu_o)}$
c $t = \frac{\lambda}{2(\mu_o - \mu_e)}$ d $t = \frac{\lambda}{2(\mu_e - \mu_o)}$

Q 15 The formula for the thickness of Half Wave Plate made up of negative crystal is

a $t = \frac{\lambda}{4(\mu_o - \mu_e)}$
c $t = \frac{\lambda}{2(\mu_o - \mu_e)}$

b $t = \frac{\lambda}{4(\mu_e - \mu_o)}$
d $t = \frac{\lambda}{2(\mu_e - \mu_o)}$

Q 16 Which of the following material is not used to polarize the light

- a Tourmaline
c Calcite

- b Polyvinyl alcohol (PVA)
d None of a, b and c

Q 17 Which of the following is not optically active?

- a Sugar solution
c Quartz

- b Cinnabar
d None of a, b and c

Q 18 Biot's law of optical activity is given by

a $\theta = S lc$

b $\theta = \frac{S}{lc}$

c $\theta = \frac{S}{l/c}$

d For a given temperature and for a given wavelength, $\theta = S lc$

Q 19 Light is passed through polarizer; the intensity remains constant on rotating the polarizer. The light is then passed through a Quarter Wave Plate and then through polarizer. The intensity changes from zero to maximum and then to zero on rotating the polarizer. This indicates that the light is

- a Plane Polarized
c Elliptically Polarized

- b Circularly Polarized
d Partially Polarized

Q 20 Light is passed through polarizer; the intensity changes from minimum to maximum and then to minimum. The light is then passed through a Quarter Wave Plate and then through polarizer. The intensity changes from zero to maximum and then to zero on rotating the polarizer. This indicates that the light is

- a Plane Polarized
c Elliptically Polarized

- b Circularly Polarized
d Partially Polarized

REFERENCE BOOKS

1. Fundamentals of Physics Extended, David Halliday, Robert Resnick, Jearl Walker,, John Wiley & Sons
2. The Feynman Lectures on Physics (3 Volume Set), by Richard Phillips Feynman (Author), Robert B. Leighton (Contributor), Matthew Sands (Contributor), The New Millennium Edition, Pearson Education India
 - Excellent websites on this book
 - v. www.feynmanlectures.caltech.edu/
 - vi. www.feynmanlectures.info/
3. A Textbook of Engineering Physics, M N Avadhanulu & P G Kshirsagar, 10th Edition, S. Chand and Company
4. Fundamentals of Optics, by Francis Jenkins, Harvey White , Tata Mcgraw Hill Publishing Co Ltd
5. Optics, Ajoy K. Ghatak, 5th Edition, McGraw Hill Education,
6. Optics, Eugene Hecht, 4th edition, Addison-Wesley
7. M. Born and E. Wolf, Principles of Optics, Cambridge University Press
8. A Text Book Of Optics, Brijlal, Dr. N. Subrahmanyam, Dr. M. N. Avadhanalu, 25th Edition, S. Chand and Company
 - i.

WORLD WIDE WEB

9. <https://www.photonics.com/>
10. SPIE - the international society for optics and photonics: spie.org/
11. Optical Society of America (OSA): <http://www.osa.org/>
12. Optical Society of India: www.osiindia.org/

CHAPTER 4

Lasers



When invented in 1960, laser was being called as “solution looking for a problem”. However, in next 50 years, several applications of laser became possible. The photograph on the left shows how laser is used in eye surgery. The photograph on the right shows 20 extremely powerful beams of neodymium glass lasers being used in a fusion reactor called as “SHIVA” - a nuclear fusion facility at Lawrence Livermore Laboratory (USA). The laser beams in SHIVA produce a power of 2×10^{10} kW in 10^{-10} s!. What is laser? And how does it have such diversified applications?

Answer to these questions is in this chapter

Index

4.1 INTRODUCTION: WHAT IS LASER?

A coherent light

4.2 CHARACTERISTICS OF LASER

What makes laser significantly applicable in technology

4.3 BASIC PHYSICS BEHIND LASER

Why laser is a synthetic light

4.4 RUBY LASER

High power and pulsed laser

4.5 HELIUM NEON LASER

Low power and continuous laser

4.6 SEMICONDUCTOR LASER

A compact and versatile laser

4.7 APPLICATIONS OF LASER

Why laser is called 'light fantastic'

4.8 FIBER OPTICS COMMUNICATION:

Laser can carry information

4.9 HOLOGRAPHY

A promise for superior information storage

4.1 INTRODUCTION: WHAT IS LASER?

A coherent light

Laser is an acronym of *Light Amplification by Stimulated Emission of Radiation*. The concept of stimulated emission was predicated by Albert Einstein in 1917. In 1954, this principle was successfully used by Charles Towns (California Institute of Technology) in constructing the first MASER (Microwave Amplification by Stimulated Emission of Radiation). In 1958, he suggested that the idea of stimulated emission could be used in visible part of the electromagnetic spectrum. Using his idea, the first laser was successfully developed by T.H. Maiman in USA. Laser is basically a coherent and amplified light. Its other features such as directionality, monochromaticity, focussibility are the consequence of its coherence. Soon after its invention, an intense research was followed due to which variety of lasers were developed. Since its invention, laser has found many applications in science, technology, industry and even day-to-day life. Indeed, laser is one of the most outstanding inventions of 20th century. This chapter is aimed at the discussion of Physics behind the laser, its design, the laser sources and applications.

Coherence in laser:

As it can be noted from Figures (4.1 to 4.2), an ordinary light is neither coherent, nor directional and nor monochromatic. A directional light may not be coherent and monochromatic, a directional and monochromatic light may not be coherent, and however, a coherent light is monochromatic, directional as well as sharply focusable

4.2: CHARACTERISTICS OF LASER:

What makes laser significantly applicable in technology

Laser is an extraordinary light having six distinct characteristics. These characteristics are mentioned below

Laser beam is highly coherent: As shown in Figure, the waves in laser are coherent. Two waves are said to be coherent, if they have zero or constant phase difference between them. Laser has two types of coherence which are discussed below

Temporal coherence: Consider Fig. If the phase difference between any two points on a wave in a direction along with the wave remains constant with respect to time, the coherence is called as temporal coherence. The length over which the phase difference remains constant is called as *coherence length*. The corresponding time is called as coherence time. The typical coherence length for a sodium light is approximately 0.3 mm, while that for a laser is about 100 m. Let ϕ_1 and ϕ_2 be the phase at point a and b at an instant t in Fig.... Let ϕ'_1 and ϕ'_2 be the phase at the same time but for another instant t' . Let $t \neq t'$. Now if $\phi_1 - \phi_2 = \phi'_1 - \phi'_2$, then the condition is called as temporal coherence.

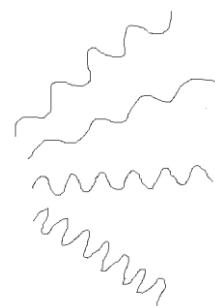


Fig 4.1 (a) Ordinary light (such as from tungsten) is neither coherent, nor directional and nor monochromatic

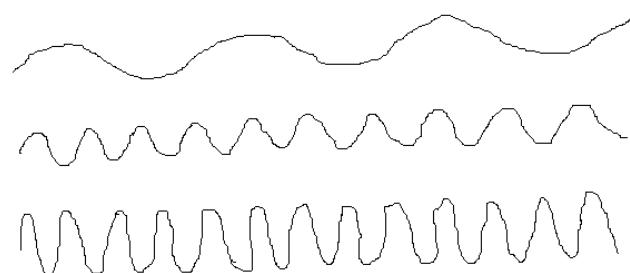


Figure 4.1 (b). A directional but polychromatc and incoherent light

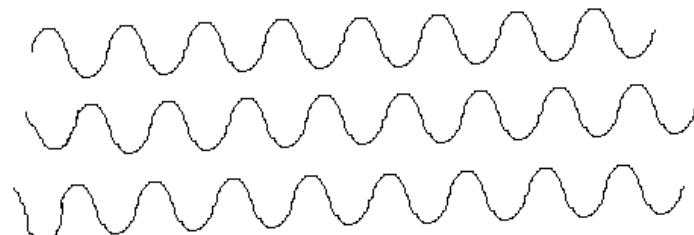


Figure 4.2 (c): Monochromatc, directional but incoherent light

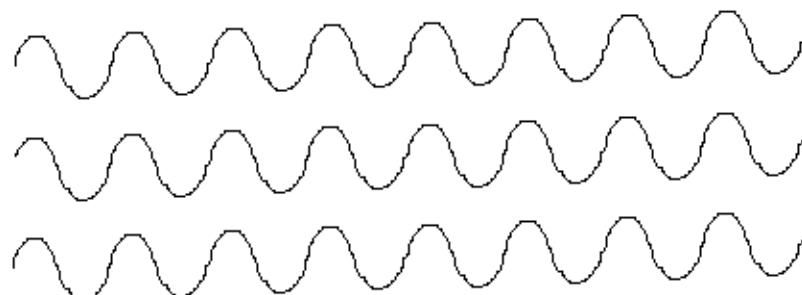


Figure 4.1 (d) Coherent (and hence directional, monochromatc and sharply focusable) light

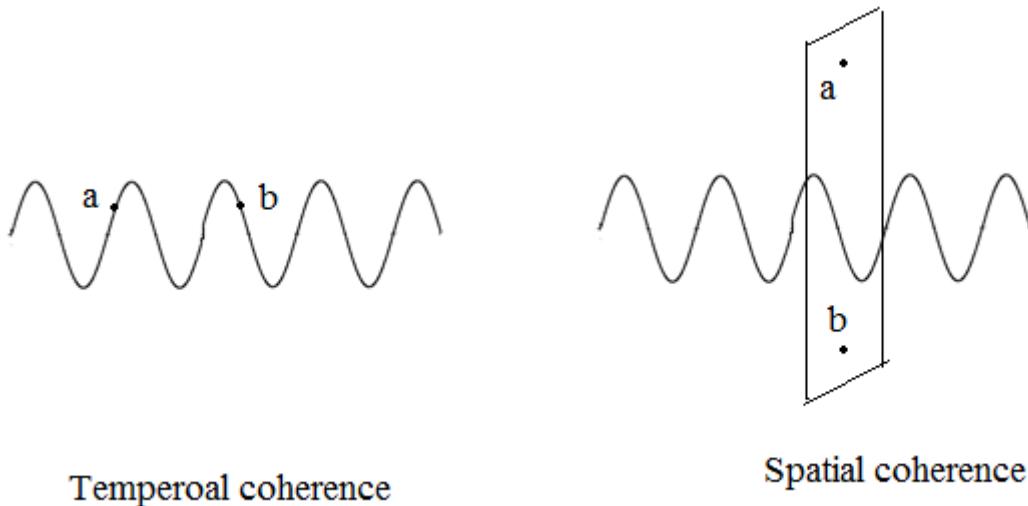


Figure (4.2) Temporal coherence (a) and spatial coherence (b)

Consider Fig. If the phase difference between the points (mentioned as a and b in Fig) laying in a plane perpendicular to the propagation of the wave remains constant with time, then it is called as spatial coherence.

As laser beam is highly coherent. Two identical but independent laser sources can generate a well-defined interference pattern

Laser light is highly powerful: Laser light is produced due to amplification of the light. Further, it is highly monochromatic and directional. Therefore it is powerful than ordinary light. Being highly directional, the power of the laser beam can be maintained over the large distances. To achieve the power equal to that of some extremely powerful laser beams, a hot object would have to be raised to a temperature of 10^{30} K.

Laser is highly monochromatic: The monochromaticity of laser is considerably high as compared to the conventional monochromatic sources. For conventional monochromatic sources the precision is up to 1 part in 10^6 , while for laser it is 1 part in 10^{15} . However, a perfect laser, which can have all wavelengths to be exactly identical, is not possible. It may have a spread of $\Delta\lambda$. The corresponding spread in frequency that is $\Delta\nu$, is called as *line width*. For white light the line width is up to 10^{14} Hz, while for laser it is 100 Hz.

Laser is highly directional: Ordinary light can be made directional with the help of a lens or mirror. However, the directionality of laser is quite high as compared to such light. The divergence of laser is in fraction of a radian. The directionality of the laser is because of the fact that in the laser cavity, the off axis photons are absorbed while only axial photons are transmitted. However, laser beam is not perfectly directional. Its divergence is extremely small but not zero. This small divergence is because of the fact that, laser is diffracted from the exit aperture of the laser cavity. The directionality of laser is so high that it has been used to measure the earth moon distance. The earth moon distance is 384,400 km. The astronauts of the Apollo mission had installed a reflector on the surface of the Moon. A laser beam was transmitted towards the moon. It was detectable even though it travelled earth-moon-earth distance.

Laser can be sharply focussed: Being highly parallel and monochromatic, laser can be sharply focussed. Energy flux densities up to 10^{16} W/cm² are possible. For ordinary light the energy flux density is about 10^3 Watts/cm². Owing to its focussibility, laser can be used as sharp knife for surgical operations. On the other hand the enormous power density due to focussing can be used for cutting, welding and drilling and in military weapon.

Laser light is polarized: This property owes to the fact that both stimulating as well as stimulated photons vibrate in the same plane.



Charles Townes(1915-2015): He was an American Nobel prize winning Physicist, who is credited for the construction of first *maser* (Microwave Amplification by Stimulated Emission of Radiation) and laying down the theoretical foundations of laser. He studied at Duke University and obtained Ph.D in California Institute of Technology. He was then appointed as a professor in Columbia University, where he invented the *maser*, the intense microwave radiation. Masers find applications in the high precision atomic clocks and in radio-wave astronomy. In 1958, he outlined the ideas behind laser in a paper published in a reputed journal. Along with the several awards and honors, he won two Nobel prizes in Physics, one in 1964 for laying down the theoretical foundations of laser and another in 1981 for precision spectroscopy using laser. In the subsequent paragraph we discuss the Physics behind the laser

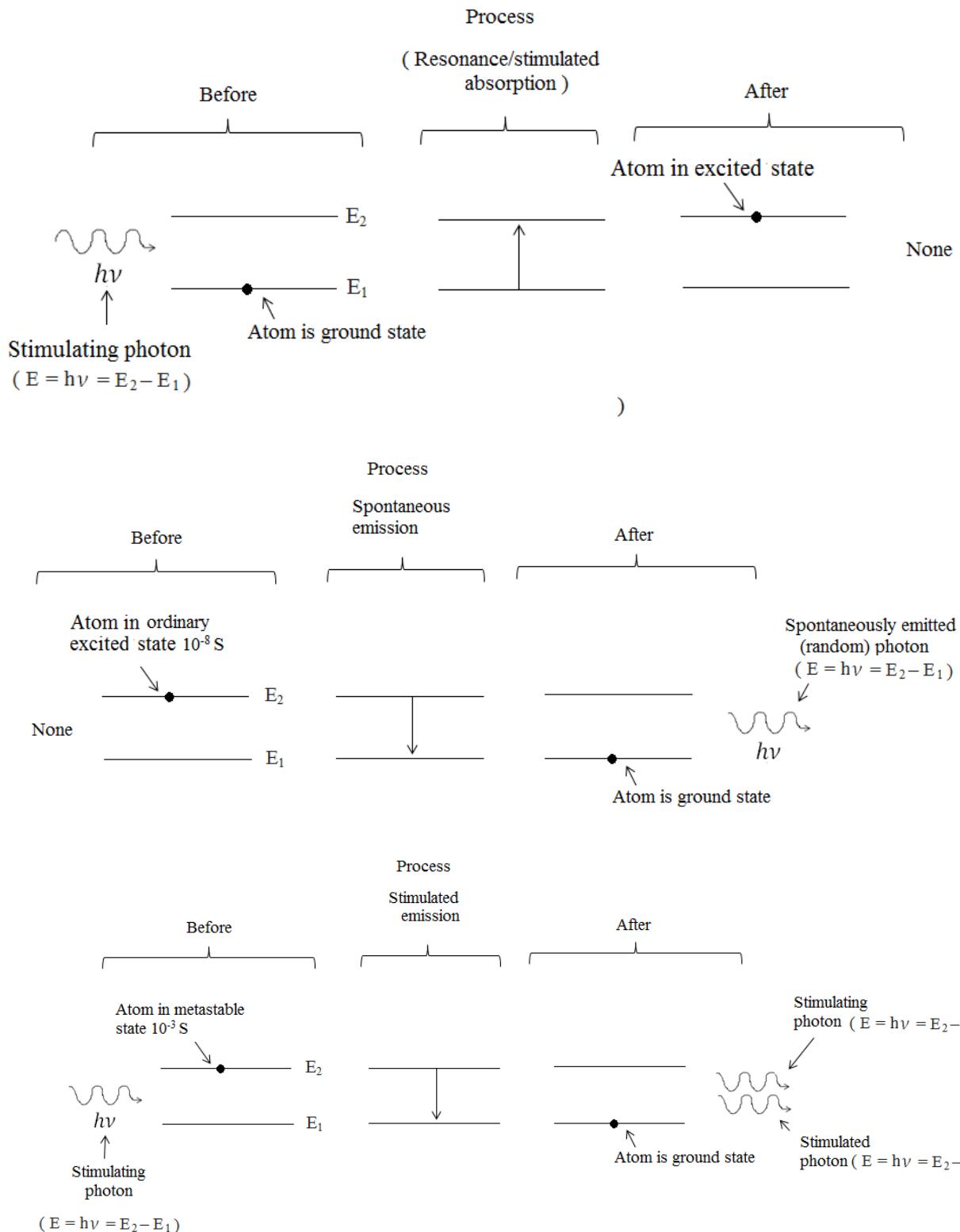
4.3 BASIC PHYSICS BEHIND LASER

Why laser is a synthetic light

For understanding the physics behind the laser, it is necessary to know the interaction between the energy (photon) and matter (atom). It is known that atoms and molecules have discrete energy levels. A photon can interact with the atom in three ways

Stimulate/resonance/induced absorption: Consider Fig (4.3a). Two energy levels of an atom; E_1 (ground state) and E_2 , excited state have been indicated. The atom is in ground state (E_1). If a photon having its energy exactly equal to $E_2 - E_1$ is incident on such atom, the photon is absorbed and the atom is raised to excited state E_2 . In this process the photon disappears, and thus the radiation is attenuated. The excitation of the atom does not take place spontaneously, it requires a photon. Therefore such absorption is also called as stimulated/induced absorption. If the photon has its energy different than $E_2 - E_1$, it is not absorbed. For absorption the photon must have its energy exactly equal to $E_2 - E_1$. Therefore this absorption is also called as resonance absorption. The number of stimulated absorptions depends upon the photon density and the number of atoms in ground state.

The stimulated absorption can be represented by following equation



**Figure (4.3) (a) Resonance/stimulated absorption (b) Spontaneous emission
(c) Stimulated emission**

$$h\nu + A = A^*$$

Spontaneous emission: Refer Fig (4.3b) Consider an atom having two energy levels E_1 (ground state) and E_2 (excited state). Let the atom be in state E_2 . The life time of the ordinary excited states is very short that is $\sim 10^{-8}$ s. The atoms stays in this state for 10^{-8} and then de-excites on its own accord to the ground state. Unlike, the stimulated absorption, which requires a photon, this process does not require a stimulating photon. Therefore this process is called as spontaneous emission. There is no external control on this process; therefore the photon is emitted in the random manner. Now consider an assembly of several atoms in the ordinary excited state. These atoms undergo spontaneous emissions. As these emissions are random, they take place at different instants and emit photons in random directions, with random phases, polarization states and with different energies and frequencies. Sometimes, all the emissions take place through the same pair of energy levels. The resulting light is monochromatic but incoherent. Such light is called as ordinary light. The number of spontaneous emissions in a given time interval depends only upon the number of atoms in excited states. The spontaneous emission can be represented by following equation

$$A^* = A + h\nu$$

Stimulated emission: Refer Fig (4.3c). Let E_1 and E_2 be two energy levels of an atom. E_1 is the ground state and E_2 is the excited state. There are two kinds of excited states. The ordinary excited states have a life time of 10^{-8} s. Some other excited states are relatively long lived. Their life time is roughly 10^{-3} s. This means that the life time of such states is one lac times greater than ordinary excited states. These states are called as *metastable states*. Let the excited atom be in metastable state. If a photon having energy exactly equal to $E = h\nu = E_2 - E_1$ is incident on an atom in metastable state E_2 , (the lower level being E_1) then it stimulates (triggers) the atom to de-excite. The photon triggering/stimulating the emission is called as stimulating photon and the photon thus emitted is called as stimulated photon. The stimulating and stimulated photons are identical in all respects. They are emitted at the same instant. Their energy, frequency, wavelength, amplitude, phase, direction and plane of vibration are same. In this process, incidence of one photon results in the emission of two exactly coherent photons. Thus this process is called as *Light Amplification by Stimulated Emission of Radiation*. The acronym of these words is LASER. In stimulated absorption, the radiation is attenuated while in stimulated emission it is enhanced/amplified. Consider a situation where, many atoms have been raised to metastable states. In such case, one stimulating photon can start the chain of stimulated emissions. (Refer Fig 4.3c). The two photons emitted due to first stimulated emission are incident on the two atoms in the metastable states, thus four identical photons are emitted. These four photons cause four more stimulated emissions, thus eight identical photons are emitted. The process thus continues and soon after first stimulated emission, the laser cavity is filled with several coherent photons. This is called as avalanche multiplication. The number of stimulated emissions taking place in a given time interval is proportional to number of atoms in metastable state as well as number of stimulating photons.

The stimulated emission can be represented by following equation

$$h\nu + A^* = A + 2h\nu$$

If atoms are in metastable states and if photon density is not enough, spontaneous emission will dominate over the stimulated emission. If the photon density is enough, but atoms are in ordinary excited states (10^{-8} s), then also spontaneous emission will dominate over the stimulated emission. The light thus generated will be ordinary. If photon density is enough and if the atoms are in ground state, stimulated absorption will dominate over stimulated emission. If the spontaneous emission dominates, then the light thus emitted will be ordinary and if stimulated absorption dominates then the light will not be amplified but will be attenuated. This indicates that for the production of laser, having enough number of atoms in metastable state as well as having enough number of stimulating photons is necessary. These conditions can be created only artificially. Therefore, laser is not a natural light, it is synthetic. Thus for laser action to be possible, it is necessary to have (a) more number of atoms in metastable state than in ground state (b) metastable state (state with relatively longer life time) (c) enough number of photons in the system

Population inversion:

In any laser system, the stimulated absorption, spontaneous emission and stimulated emission take place simultaneously. As discussed in previous sections, if the number of atoms in lower energy level is more than that in metastable state, then stimulated absorption will dominate over the stimulated emission. As a result, the light will be attenuated rather than getting amplified. Thus we need to have more number of atoms in metastable state than in ground state.

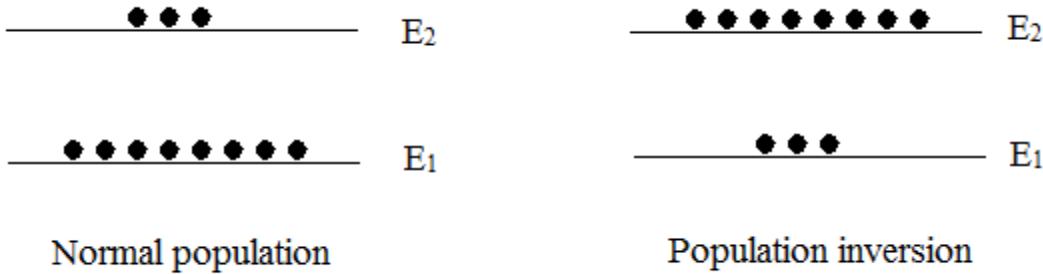


Figure (4.4) Concept of Population Inversion

According to Boltzmann, the number of atoms (N) per unit volume in an energy level E is given by

$$N = N_0 e^{-\frac{E}{kT}}$$

Where N_0 is a constant, k is Boltzmann constant and T is the temperature.

The above equation clearly indicates that at any temperature, small or large, the number of atoms in higher energy level is lesser than that in the lower energy level. Consider two energy levels E_1 and E_2 . Let the number of atoms in these levels be N_1 and N_2 respectively. We can then write

$$N_1 = N_0 e^{-\frac{E_1}{kT}} \text{ and } N_2 = N_0 e^{-\frac{E_2}{kT}}$$

The ratio of N_1/N_2 can be written as

$$\frac{N_1}{N_2} = N_o e^{-\frac{(E_1-E_2)}{kT}} \quad \dots (4.1)$$

As $E_2 > E_1$, we have $N_1 > N_2$. Thus, at any finite and positive temperature, the number of atoms in the higher energy level is always less than that in lower energy level. This is equilibrium condition. However, for stimulated emission to dominate over stimulated absorption, we require $N_2 > N_1$. This is called as population inversion. The equation ... indicates that this is possible only when the temperature is negative. Therefore the state of population inversion is called as *negative temperature state*. However, it may be noted that the physically the temperature is not negative. The word negative temperature state indicates a non-equilibrium situation. As population inversion is a non-equilibrium state, clever techniques have to be used to achieve it. This is called as pumping

Pumping: As it can be noted from the previous section, clever techniques are necessary to have more number of atoms in metastable state than in ground state. These techniques involve providing the energy in appropriate form to the system of atoms. Due to provision of energy, the atoms are raised to metastable states. As the excited state is metastable, the atoms stay there for relatively longer time. The process of provision of energy to achieve population inversion is called as *pumping*. Depending upon the type of laser, pumping can be done in the variety of ways. The form of energy required for pumping can be optical (flooding a powerful light), electrical (discharge of electricity) or forward biasing the PN junction (diode laser). Optical pumping is generally used for solid state lasers, in which the active system is transparent to the light. Electrical pumping is used in case of gas lasers. The other techniques of pumping involve chemical pumping, nuclear pumping, X ray pumping etc.

Metastable states: As discussed earlier, if the atoms are pumped in ordinary excited states (10^{-8} s), then they will not stay there for a longer time. Therefore there will be great chances of spontaneous emission. This because the atoms will not 'wait' for the stimulating photons to be incident. Similarly if the pumping is performed to ordinary states, then population inversion cannot be maintained as the atoms will keep on coming to ground states within 10^{-8} s. Therefore the states whose life time is considerably higher than the ordinary states are necessary. For some atomic and molecular species, such states are available. The life time of these states is roughly 10^{-3} s. This means that such states are one lac times longer than the ordinary states. Thus atoms 'wait' for sufficient time for the stimulating photons. Such states are called as *metastable states*. It is to be noted that population inversion as well as stimulated emission is not possible without metastable states

Active system: An atomic or molecular system with metastable state and having population inversion is called as active system. Active system is characterized by optical gain (amplification). The active system can be in solid, liquid or gas. It may be noted that all types of atoms are not suitable for laser action. In most of the active media, there are few active centres. The rest of the medium serves as host for active centres.

Resonant cavity: As discussed earlier, for laser action, metastable state and population inversion is not sufficient. It is equally necessary to have enough number of stimulating photons in the system. This can be achieved by placing the active medium in between two mirrors. Without such mirrors, the photons produced due to stimulated emission would escape from the laser cavity. Such photons will not be available for producing the stimulated emissions of the remaining atoms in metastable state. Such atoms may de-excite spontaneously. Thus it is necessary to trap as many stimulating photons as possible in the laser cavity. This purpose is served by the two mirrors across the laser cavity. One of these mirrors is 100 % reflecting and another is partially reflecting. The distance between these two mirrors is $d = n \frac{\lambda}{2}$, where λ is the wavelength of the laser being produced. Due to such distance, the laser cavity resonates at λ , due to which the laser having wavelength λ is enhanced. Due to this reason, such cavity is called *resonant cavity*. Resonance cavity plays many roles, which are discussed below

- a. Due to $d = n \frac{\lambda}{2}$, the resonant cavity enhances the laser having wavelength λ . The other wavelengths, which do not resonate, are suppressed. Thus the laser becomes perfectly monochromatic.
- b. The resonant cavity supports positive feedback. We know that, in electronics, if the amplifier is provided with positive feedback, then it oscillates indefinitely. The mirrors used across the resonant cavity reflect back the photons produced due to stimulated emissions. Such photons when incident on the atoms in the metastable state, produce more number of photons. Thus the mirrors help in providing the positive feedback of the photons to the laser cavity. It may be noted that, in the beginning, there are no enough number of photons. Thus the first emission is always a spontaneous emission. However, photons produced due to first few spontaneous emissions trigger the stimulated emissions. As discussed earlier, this begins the process of avalanche multiplication of laser photons. Due to positive feedback, the laser oscillates in the resonant cavity. When the laser acquires sufficient intensity, it bursts out from the partially reflecting mirror. It may be noted that if both the mirrors are 100 % reflecting then the laser will oscillate but will never come out.
- c. We know that the stimulating photon imposes all its properties upon the stimulated photon. The directions of the stimulating and stimulated photons are same. The mirrors exist only across the laser cavity. Due to back and forth reflections of the laser photons, the beam along the axis of the laser cavity is built up. The photons travelling off axis escape from the cavity (refer Fig), thus the laser along the axis builds up and thus becomes more and more directional. Thus the resonant cavity, due to presence of the mirrors makes the laser highly directional. It may be noted that in absence of the mirrors, there is neither a positive feedback nor a directionality in the laser beam.

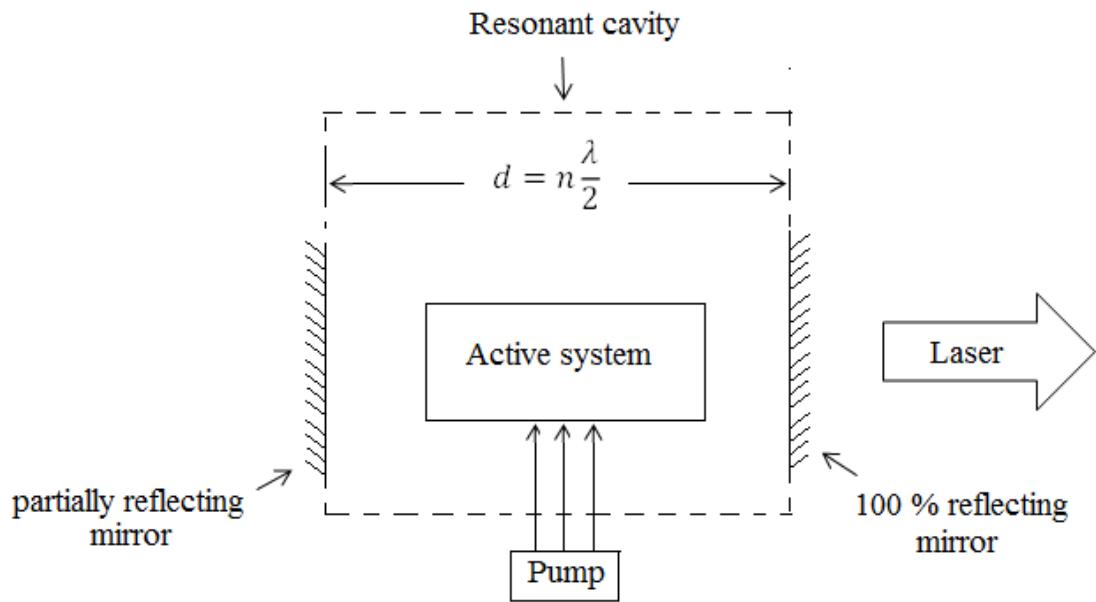


Figure (4.5) Basic components of laser

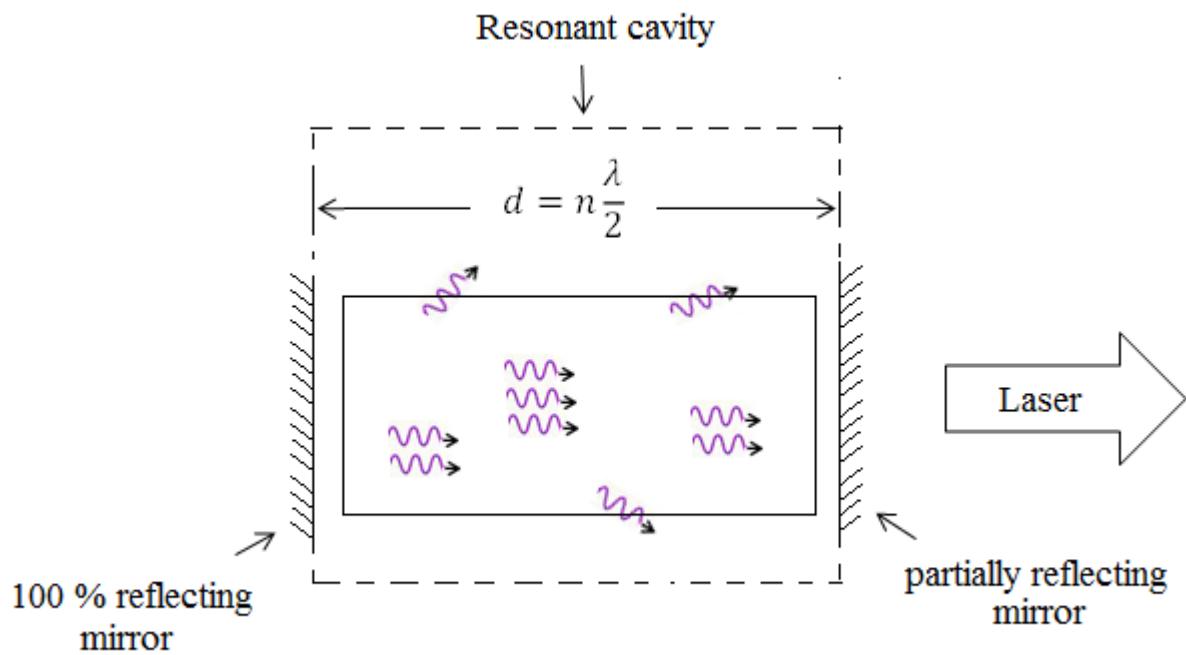
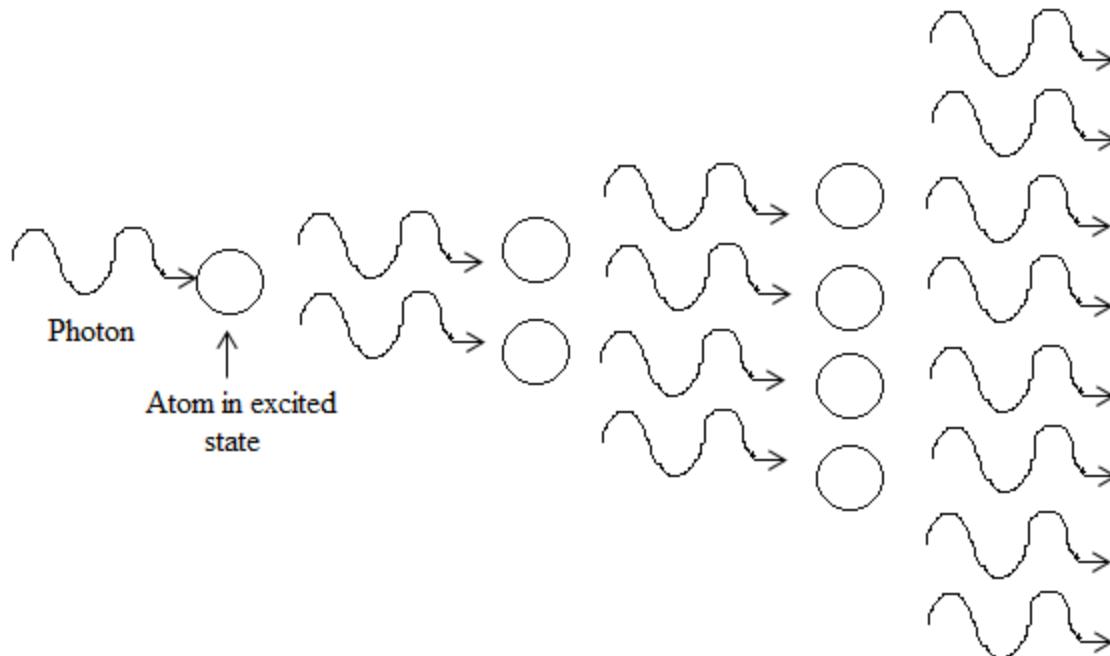


Figure (4.6) Cavity oscillations (lasing)

Cavity oscillations: These are the oscillations of the laser photons across the mirrors. As discussed earlier, the cavity oscillations support the positive feedback and provide the directional and monochromatic properties to the laser beam.

Lasing: Pumping, achievement of population inversion, the first few spontaneous or chance photons, stimulated emissions, absorption of the off axis photons, the back and forth oscillations of the laser photons within the laser cavity, the chain reaction of stimulated emissions, avalanche multiplications of the photons and outburst of the laser through the partially reflecting mirror entirely constitute a phenomenon collectively called lasing. Note that lasing is not a dictionary word. Lasing indicates the set of processes responsible for production of laser.



Figure(4.7): Avalanche multiplication of photons

Principal laser schemes(Optional): Though atoms and molecules have several discrete energy levels only few of them are utilized for lasing. Accordingly there are three types of energy level schemes used for lasing. They are, two level scheme, three level scheme and four level scheme

Two level scheme: In this scheme, only two energy levels are used for lasing. (Refer Fig) The atoms are pumped from E_1 to E_2 . Stimulated emissions take place from E_2 to E_1 . Thus the energy levels used for pumping and stimulated emission are same.

The only laser in which two level scheme is used is diode laser. Valence band and conduction band are used for lasing. Two level scheme is generally not employed for any other type of laser due to its drawback. It may be noted that, the photon which is used for pumping the atom from

E_1 to E_2 itself can cause the stimulated emission from E_2 to E_1 . Thus the chances of pumping and stimulated emission are equal. As a result, the number of atoms in E_2 and E_1 are almost same. Thus population inversion and the thus lasing is not possible.

Three level scheme: This scheme involves three energy levels, one of which is metastable. Consider Fig (4.8). The three levels involved in lasing are E_1 (ground state), E_2 (metastable state/upper lasing level) and E_3 (ordinary state: pump state). Pumping radiation having photon energy ($h\nu'$), usually light from the flash lamp is allowed to fall on the active system. The atoms are excited from E_1 to E_3 . E_3 being an ordinary state, the atoms stay there only for a short time. Then they de-excite to E_2 by emitting a photon energy ($h\nu''$). These transitions are rapid transitions. As the energy emitted is weak, these transitions are also called as *radiation-less transitions*. E_2 state is a metastable state and therefore the atoms stay there for longer time. Thus the number of atoms in E_2 state builds up and population inversion is achieved. The first transition from E_2 to E_1 produces a spontaneously emitted chance photon. As discussed, this photon triggers a chain of stimulated emissions ($h\nu$) and thus laser is produced. The three level scheme is better than two level scheme because the energy levels involved in pumping (E_1 to E_3) are different than those involved in lasing (E_2 to E_1). However this scheme requires large pumping power. Further when all the atoms lase from E_2 to E_1 , E_2 depopulates and E_1 populates. Thus the lasing temporarily ceases. It is necessary to pump the atoms once again from E_1 to E_2 . Therefore all lasers which result from three level scheme are essentially the *pulsed lasers*. The prime example of three level scheme is ruby laser.

Four level scheme: Consider Fig...which shows four level scheme. There are four energy levels involved. E_1 is ground state and E_2 , E_3 , E_4 are excited states. Amongst these excited states, E_3 is metastable and E_2 and E_4 are ordinary excited states. The atoms are pumped from E_1 to E_4 . The atoms then de-excite spontaneously from E_4 to E_3 . E_3 is a metastable state. The atoms thus stay in E_3 for longer time and therefore population inversion is achieved. The lasing takes place from E_3

to E_2 . E_2 being an ordinary excited state, the atoms spontaneously and quickly return to the ground state E_1 . Thus the level E_1 always contains enough number of atoms for pumping. Therefore pumping requires less power. The lasing (E_3 to E_2) and pumping (E_1 to E_2) take place simultaneously. Therefore the lasers produced from four level scheme are continuous. The examples of four level scheme are He Ne laser and Nd-YAG laser and CO₂ laser.

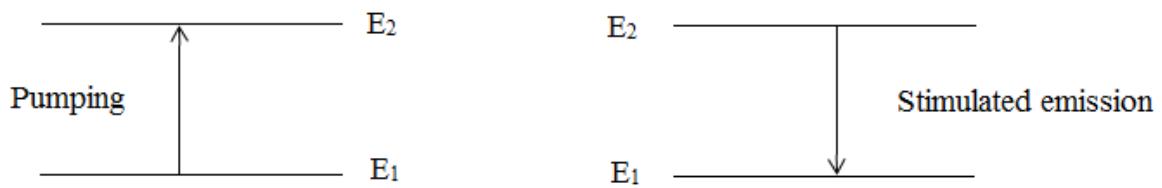


Figure (4.7): Two level Scheme

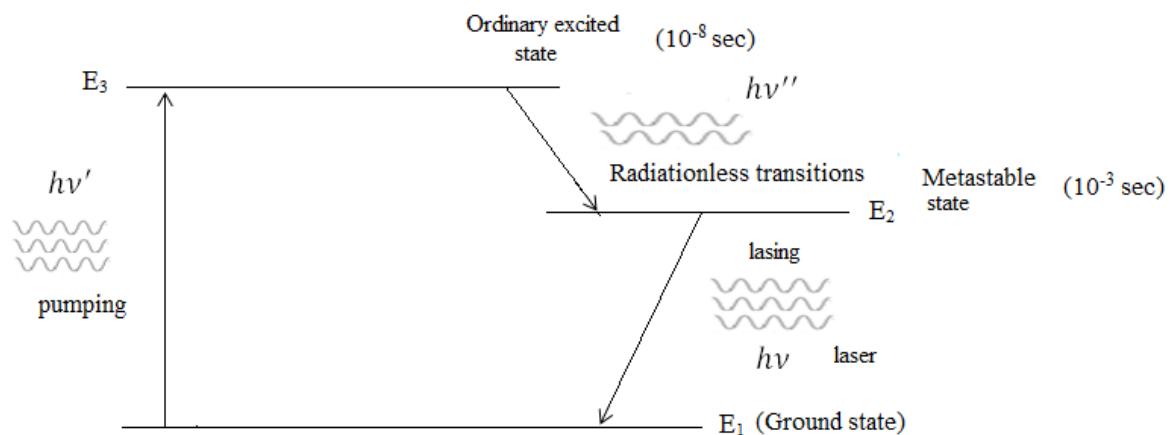


Figure (4.8) Three level scheme

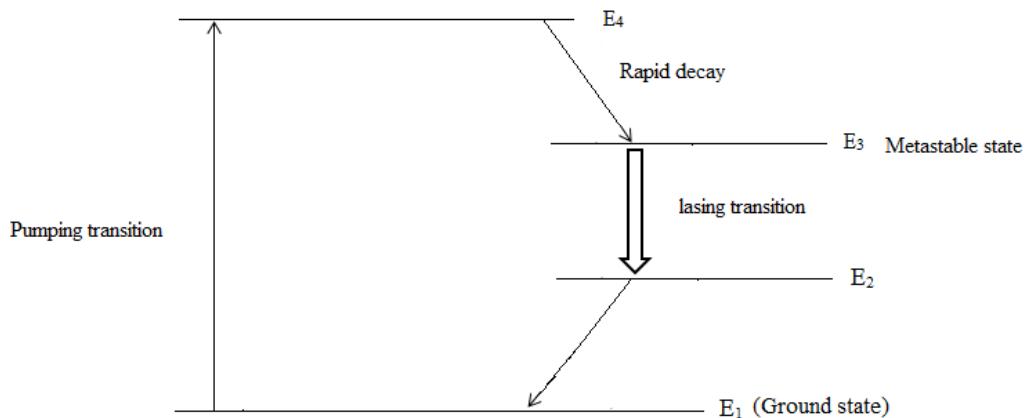
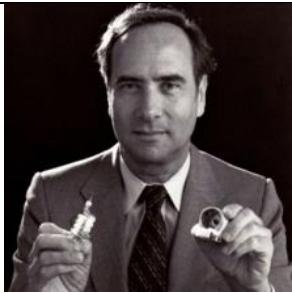


Figure (4.9) Four level scheme



Theodore H. Maiman(1927-2007): He was an American Physicist who is credited for the invention of the world's first ever laser, the Ruby laser. He graduated in University of Colorado and obtained his post graduate degree and Ph.D. in Stanford University. He then joined Hughes Research Laboratories, where he invented the ruby laser. The idea behind this work was proposed by Charles Towns in 1958 and since then several research groups, including those at IBM, Bell Labs, MIT (Boston) and Columbia University were pursuing the Town's suggestion. However Maiman was the first to realize the idea in practice. Maiman published his invention in 'Nature' and was also awarded a patent for this invention. Later on he earned many patents and won many awards and honors. Time magazine cited Maiman's invention of the laser as among the twenty most important technological developments of the 20th century. In the subsequent paragraphs, we discuss Ruby laser the invention of which led to the development of the other lasers.

(4.4) RUBY LASER: (Optional)

High power and pulsed laser

Ruby laser is historically the first laser which was constructed and operated by T.H. Maiman in Hughes Research Laboratories, USA in 1960. This invention was based on a suggestion of Charles Towns in 1958. Ruby laser is a solid state laser. Ruby crystal consists of Al_2O_3 in which a few Al^{+3} ions are replaced by Cr^{+3} ions. The doping level of Cr^{+3} ions is 0.05 %. The active system in ruby laser is Cr^{+3} ions, while Aluminum and Oxygen are inert and they serve as host. Cr^{+3} ions strongly absorb blue and green color due to which ruby appears pinkish.

Construction: A ruby crystal in the form of cylindrical rod, having length 4 cm and diameter 1 cm is surrounded by a helical Xenon flash tube. The end surfaces of ruby rod are grounded, polished and silvered. The silver coating is such that the rarer surface works as fully reflecting mirror and the front surface works as semi-reflecting mirror. The partially reflecting front surface works as a window for the laser output. The Xenon flash tube is used for optical pumping. It is connected with high tension power supply. It emits a powerful beam of white light. A small fraction of this light is used to excite the Cr^{+3} ions while the rest is converted into heat. As the intense heat is undesirable for the ruby crystal, cooling system based on liquid nitrogen is necessary.

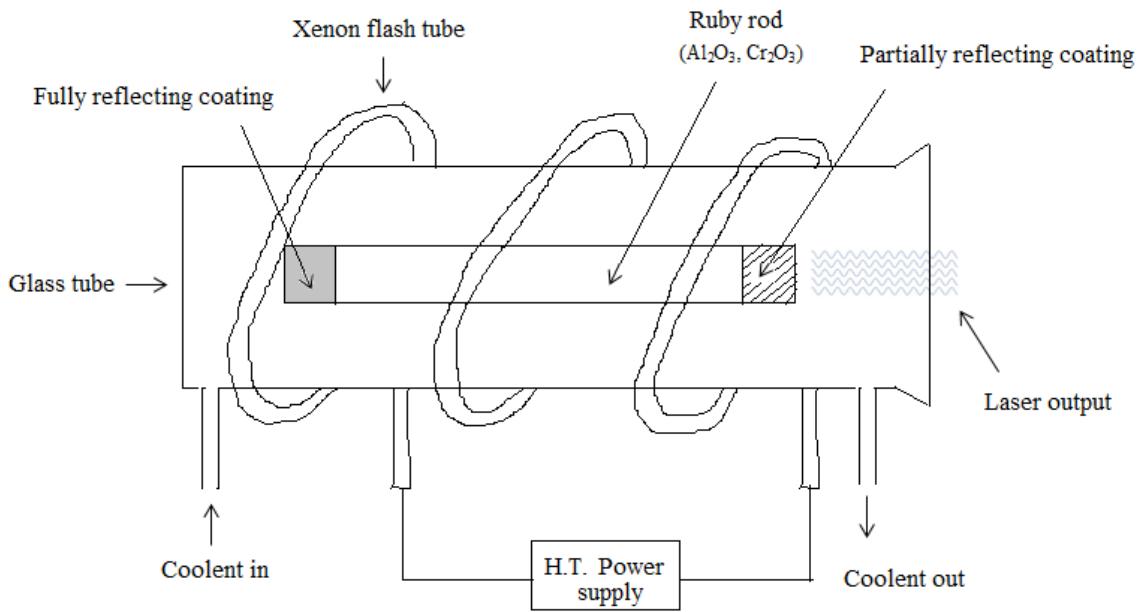


Figure (4.10) : Schematic diagram of Ruby laser

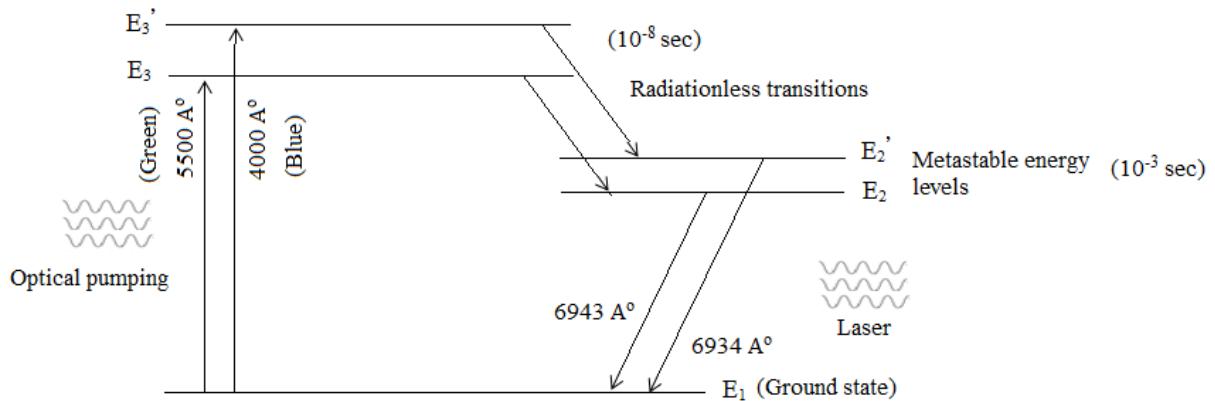
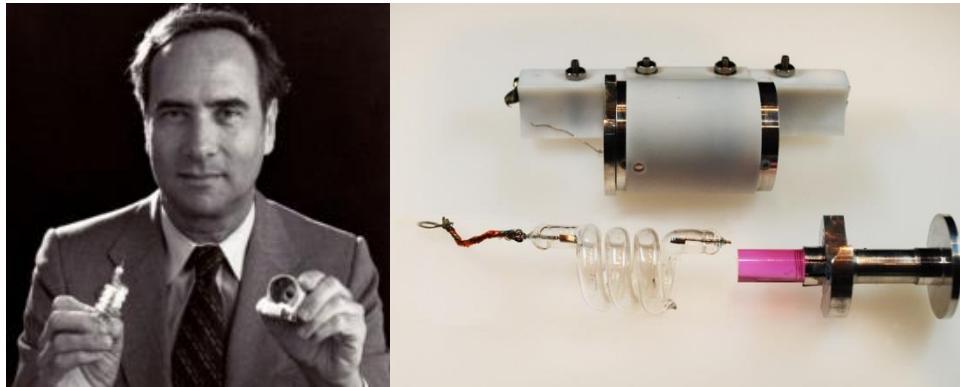


Figure (4.11) Ruby laser: Energy level diagram of Cr^{+3} ions

The ruby rod is a lasing medium and its length is such that it works as optical resonator. Due to the mirrors on the either side of the ruby rod, the photons parallel to the axis of the rod are sustained while the off axis photons are absorbed.

Working: Refer the energy level diagram of Cr^{+3} ions as shown in Fig (4.11). The white light emitted by the xenon flash tube is incident on the ruby rod. The blue and green part of the light having wavelengths 4000 \AA and 5500 \AA is absorbed by the Cr^{+3} ions due to which they are

pumped from ground state (E_1) to E_3 and E_3' energy levels respectively. The E_3 and E_3' levels are ordinary



Theodore Maiman with his Ruby laser

Components of Ruby laser, the pink ruby rod can be clearly seen

levels having a life time of about 10^{-8} sec. After the life time, a few Cr^{+3} ions de-excite to E_1 while rest of the ions de-excite to E_2 and E_2' . The energy emitted in these transitions is small and is utilized for the thermal vibrations of the atoms. Therefore these transitions are also called as radiationless transitions. Both E_2 and E_2' levels are metastable having a life time of about 10^{-3} sec. Thus now majority of the Cr^{+3} ions are in the metastable E_2 and E_2' states while only a small fraction of the ions is E_1 state. Thus the required population inversion is achieved. The first photon emitted due to de-excitation of Cr^{+3} ions is spontaneously emitted and is called as chance photon. This photon triggers the lasing of excited Cr^{+3} ions. When this photon falls on another excited Cr^{+3} ion, there occurs a stimulated emission due to which two photons are emitted. As discussed earlier, these two photons are identical in all the respects such as phase, energy, wavelength, frequency, and direction. These two photons are incident on another two Cr^{+3} ions, due to which four identical photons are produced. Subsequently there occurs an avalanche of stimulated emissions of rest of the Cr^{+3} ions. The off-axis photons are absorbed while the photons along the axis move back and forth between the end mirrors. The interaction of these photons with excited Cr^{+3} ions results in the amplification of the light and thus a highly coherent beam of laser is produced. When the laser grows in sufficient intensity, it bursts out from the semi-reflecting mirror. The transition from E_2' to E_1 results in the emission of 6934 \AA while the transition from E_2 to E_1 results in the emission of dominant 6943 \AA which corresponds to red color. The group of atoms collected in E_2 and E_2' results in an intense pulse of ruby laser. When all the atoms collect in E_1 , the next flash of xenon tube is operated. Thus ruby laser is a pulsed laser.

As discussed earlier, only a small fraction of the flash is used for the excitation of Cr^{+3} ions, the rest is converted into heat. A coolant (liquid nitrogen) is necessary for the dissipation of the heat otherwise the heat not only may damage the ruby crystal, but may also cause the thermal vibrations of the atoms resulting in noise. Ruby laser is operated in pulses due to its high power.

The efficiency of ruby is less than the other lasers. It is used for holography, ranging, drilling holes in diamonds and scientific research.



Ali Javan (1926- 2016): He obtained his education from Columbia University. His thesis advisor was Charles Towns. He then joined Bell Telephone Laboratories where he designed and fabricated the first gas laser. In 1960 he joined MIT, Boston and remained there as Professor emeritus. His other contributions in Physics are atomic clocks, optical antenna for emitting and receiving light, accurate measurement of speed of light etc

4.5 HELIUM NEON LASER:(Compulsory)

Low power and continuous laser

He Ne laser is a gas laser. It was invented by Ali Javan, Bennet and Harriot in 1961 at Bell laboratories. The other gas lasers are CO₂ and argon laser. He Ne laser is a low power and continuous laser. The energy levels of gas laser are more precisely defined as compared to those of solids. This has two consequences. One is that the transitions from higher to lower levels are sharper and hence the wavelengths of gas lasers are more precisely defined. However as the energy levels are sharper, optical pumping based on flash tubes is avoided. Instead, electrical discharge is preferred. For discharging the electricity through the gas, high tension power supply is necessary. Gas lasers are most widely used lasers.

Construction: Refer Fig (4.12). He Ne laser consists of a glass tube having length 10-100 cm and diameter 2-8 mm. It is filled with the mixture of Helium and Neon gas in the ratio 85 % to 15% respectively. The active system required lasing is Neon. The energy level diagram of Neon is more suitable for lasing action. Helium is used as a pumping agent. The pressure of the Helium gas is 1 mm of mercury and that of Neon gas is 0.1 mm of Mercury. Cathode and Anodes are inserted in the tube for discharge of electricity. They are connected to a High Tension power supply (10 KV) . At the ends of the discharge tube, two mirrors are fixed. One mirror is fully reflecting and another is partially reflecting. Before mirrors, Brewster windows are used for polarizing the laser. The distance between the mirrors is adjusted to $n\frac{\lambda}{2}$ where λ is the wavelength of He Ne laser

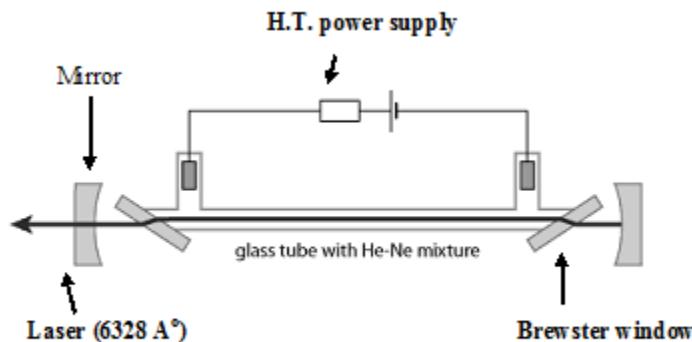
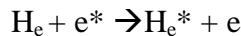
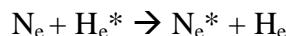


Figure (4.12): Construction of He Ne laser

Working: Fig (4.13) shows the energy level diagrams of Helium and Neon atoms. The energy levels E_2 and E_3 of Helium atoms are metastable. When H.T. power supply is made ON, the electrons are released from the atoms. These electrons are energized due to P.D. across cathode and anode. The excited electrons collide with Helium and Neon atoms. As the percentage of Helium atoms is more as compared to Neon and as Helium atoms are lighter than Neon atoms, the He atoms are more readily excited than Neon atoms. The electrons preferably excite Helium atoms from the ground state E_1 to E_2 and E_3 . This can be represented using following equation



E_2 and E_3 levels are metastable. The excited Helium atoms collide with the Neon atoms. The energy levels E_4' and E_6' of Neon atoms are close to the levels E_2 and E_3 of Helium. Consequently, the inelastic collisions result into excitation of Neon atoms from E_1' to E_4' and E_6' . This can be represented using following equation.



The pressure of Neon atoms is less as compared to Helium atoms and as the percentage of Neon atoms is quite less as compared to Helium atoms. Hence the probability of reverse excitation of Helium atoms due to collisions with Neon atoms is very rare. The energy levels E_4' and E_6' of Neon atoms are metastable. Due to continuous excitation, these energy levels are more populated than the lower levels. Thus, required population inversion is achieved. Following de-excitations are possible

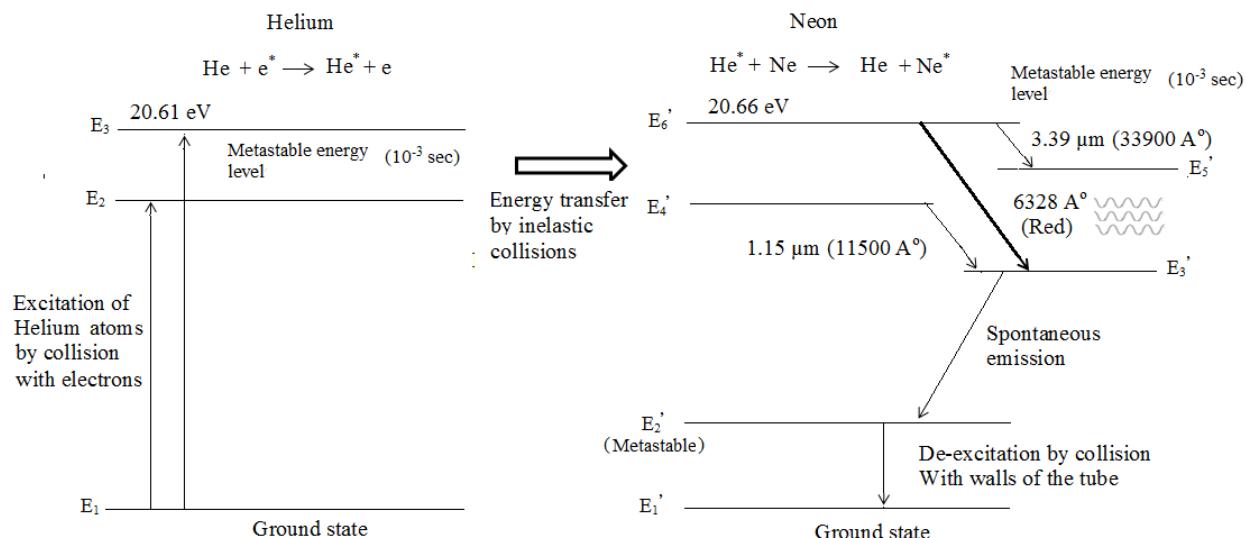


Figure (4.13): Energy level diagram of He Ne laser

E_6' to E_3' (6328 \AA° : Red color)

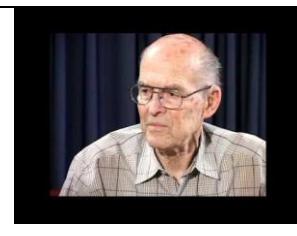
E_6' to E_5' (33900 \AA° that is $3.39 \mu\text{m}$: Infrared)

E_4' to E_3' (11500 \AA° that is $1.15 \mu\text{m}$: Infrared)

Out of above transitions, the He Ne laser is commonly operated for E_6' to E_3' (6328 \AA° : Red color). The other two wavelengths, namely E_6' to E_5' ($33900 \text{ \AA}^{\circ}$ that is $3.39 \mu\text{m}$: Infrared) and E_4' to E_3' ($11500 \text{ \AA}^{\circ}$ that is $1.15 \mu\text{m}$: Infrared) are suppressed by using highly reflective coatings at 6328 \AA° that absorb the other wavelengths. Another method for suppressing these wavelengths is to select the length of laser cavity such that 6328 \AA° dominates over the other wavelengths. As discussed earlier, the first photon is spontaneously emitted. This photon initiates the chain of stimulated emissions. These photons move back and forth between the mirrors and amplify the laser. When the laser acquires sufficient intensity, it comes out from the semi-reflecting mirror.

After these transitions, there occur spontaneous de-excitations from E_3' to E_2' . Thus the population of E_2' level thus increases. This is undesirable as E_2' is metastable level. If the atoms stay in E_2' level for longer time, there is a chance of re-excitation of the Neon atoms from E_2' to E_3' due to photons emitted because of E_3' to E_2' transitions. Secondly, the population of E_1' level decreases. This prevents the pumping of the Neon atoms from E_1' to E_4' and E_6' due to which lasing may cease. To depopulate the E_2' level, the diameter of the glass tube is made narrow. The Neon atoms de-excite from E_2' to E_1' level due to collisions of Neon atoms with the walls of the tube. E_2' level is thus depopulated. As the Neon atoms are continuously pumped from E_1' to E_4' and E_6' , He Ne laser is a continuous laser.

In certain respects, He Ne laser is better than Ruby laser. As noted earlier, Ruby laser is a pulsed laser, while He Ne laser is continuous. As He Ne laser is a gas laser, its spectral width is narrower than Ruby laser. The output of Ruby laser may be affected due to crystalline imperfections and thermal distortions. These problems are not encountered in He Ne laser. Further, unlike Ruby laser, He Ne laser does not require cooling arrangements. However, the efficiency of He Ne laser is low. Typically He Ne laser is operated at 2500 volts and few mA, while the output of laser is in mW. The applications of He Ne laser include student laboratories, holography, interferometry, barcode scanners and research.



Robert N. Hall (1919 - 2016): He was a physicist who invented semiconductor laser and received patent for it. He also invented magnetron which is commonly used in microwave ovens. He received his degree and Ph.D. from California Institute of Technology. He then joined General Electric company where he invented the semiconductor laser in 1962. In his career he received 43 patents

(4.6) SEMICONDUCTOR LASER:Compulsory

A compact and versatile laser

Semiconductor laser is a special case of LED (Light Emitting Diode). LED is a device that converts electricity into light. When a PN junction is forward biased, the electrons and holes recombine. On account of this recombination, a photon is emitted. For compound semiconductors, the photon lies in the visible region. Depending upon the energy gap of the compound semiconductor, LEDs of variety of colors such as red, green and blue and even white are possible.

While LEDs are moderately doped and operated with moderate current, semiconductor laser is heavily doped and operated with large forward current above the threshold value. Semiconductor lasers are extremely compact. Refer Fig.(4.16)and (4.17).Two side surfaces are made flat and polished. One out of these surfaces is fully reflecting whilethe other is semi-reflecting. The region in between works as a FabryParot resonator. The other two sides are roughened to avoid lasing action in that direction. Due to heavy doping, there is a

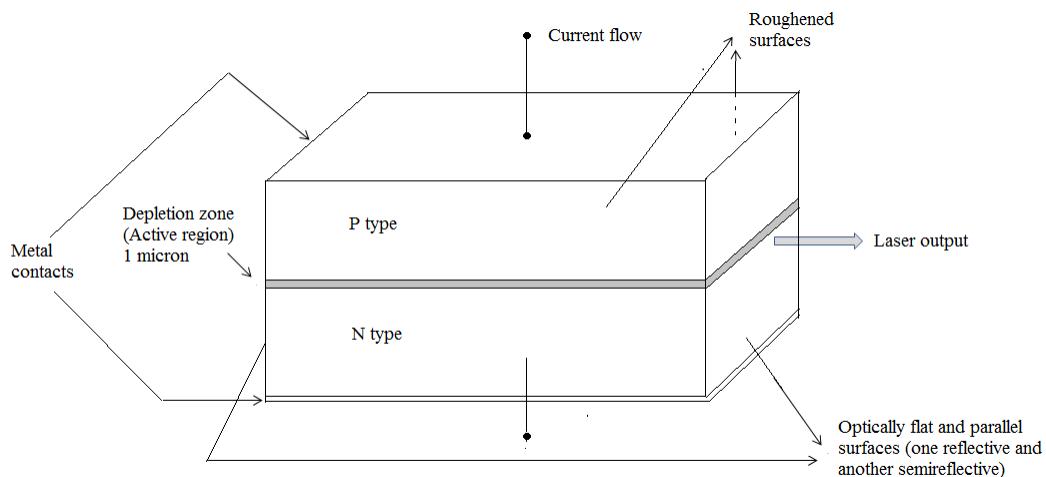


Figure (4.16): Construction of Semiconductor laser

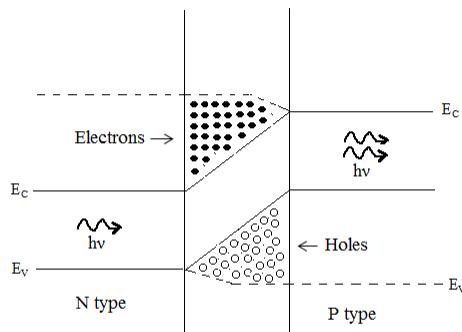


Figure (4.17) Working of Semiconducting laser.

large concentration of electrons in the conduction band of n region and large concentration of holes in the valance band of P side. If this junction is operated with a low current, then electrons and holes recombine and photons are spontaneously emitted. This results in un-coherent light. However, if large forward current above threshold is injected then heavy concentration of

electrons and holes quickly approach towards depletion zone. As shown in Fig , this results in to large concentration of filled levels of electrons in the conduction band near the junction and large concentration of vacant energy levels corresponding to existence of holes in the valance band near the junction. This results in to population inversion. Thus the heavy forward current acts as a pumping agent. The electrons and holes recombine at a fast rate due to their large concentration and heavy current. This results in to sudden production of large number of photons. These photons work as a stimulating photons for the subsequent de-excitations. This process is supported by the fully reflecting and semi-reflecting mirrors on the other side. In this manner an avalanche of photons constituting a laser is produced. If the semiconducting material is the GaAs, then the resulting light has a wavelength of 9000 \AA° (IR region), while for GaAsP, a red light of wavelength 6500 \AA° results.

As compared to the other lasers, the semiconductor lasers are advantageous in many respects. They are simple, cheap, compact, highly efficient, low power consuming and requiring less auxiliary equipment. Their only disadvantage is that they have more divergence (5° to 15°). Semiconductor lasers have revolutionized Information Technology. They are used to read and write CD, in laser printers, laser copiers and Fiber Optics Communication.

(4.7) APPLICATIONS OF LASER

Why laser is called ‘light fantastic’

Laser has six characteristic features. It is a highly coherent, monochromatic, directional, powerful, focusable and polarized light. Laser has found many applications in science, technology, industry, research and day to day life. Since its invention, varieties of lasers have been synthesized. The available lasers range from gas lasers (such as He Ne, CO₂, Argon), semiconductor lasers (GaAs, AlGaAs, GaInAsP etc.), Solid state lasers (Ruby, Nd:YAG, Nd: Glass), to Dye, Chemical and Excimerlasers. Depending upon its type, the size of a laser varies from pinhead to a room. The power of laser varies from milliwatts to 10^{13} watts. Laser can be continuous as well as pulsed (pulse duration: 10^{-9} s). Lasers find applications in diversified fields. They are used in mechanical industry, electronics, communication, eye treatment, drilling tiny holes in diamond, cutting clothes, precision surveying, precision length measurements using interferometry, generation of holograms, entertainment, communication, defense, CD players, computers, laser printers etc. A few prominent applications of laser have been discussed below.

The applications of lasers in Information technology involve holographic data storage, fiber optic communication, reading and writing a CD and laser printers



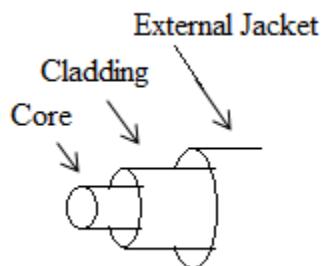
Sir Charles Kao (1933-): He is Chinese Physicist who is known as the *father of fiber optics*. He graduated from University of Greenwich. He then pursued research and received his Ph.D. degree in 1965 from University College London. In the 1960s, at Standard Telecommunication Laboratories (STL) Kao and his co-workers did their pioneering work in the realization of fiber optics as a telecommunications medium. Kao has published more than 100 papers and was granted over 30 patents all related to Fiber Optics. Along with many awards and honors, the most prestigious is Nobel prize in Physics, 2009. In the subsequent sections, we discuss Fiber optics communication...a communication using light which is replacing the conventional copper cable based communication

(4.8) FIBER OPTICS COMMUNICATION: (Optional)

Laser can carry information

Information traffic is increasing at a faster rate. The number of channels, internet connections and telephone connections are increasing day by day. Radio waves and microwaves cannot cope up with increased demand of the communication traffic. Therefore an additional agency is necessary. Light, due to its higher frequency can cope up with these demands. The capacity of the carrier for the communication that is number of bits per second depends upon it's bandwidth that is frequency. The frequency of the light is 10^{15} Hz while the frequency of the radio waves and microwaves are 10^6 Hz and 10^9 Hz respectively. Thus capacity of light for the communication is considerably higher as compared to radio waves and microwaves. After the invention of it's in 1960, laser was successfully used for the communication in 1965.

Unlike radio waves and microwaves, light cannot be used for communication in the open space, due to scattering problems. Therefore it needs to be guided through optical fibers. The materials used for optical fibers are glass and plastic. These are transparent to laser. However, if the Fiber is bent, a mechanism for guiding the laser through the bent fiber is needed. This mechanism is based on *total internal reflection*. Refer Fig The core has greater refractive index (μ_1) than cladding (μ_2). Thus if the light passing from optically denser core to rarer



Figure(4.18) Structure of Optical Fiber

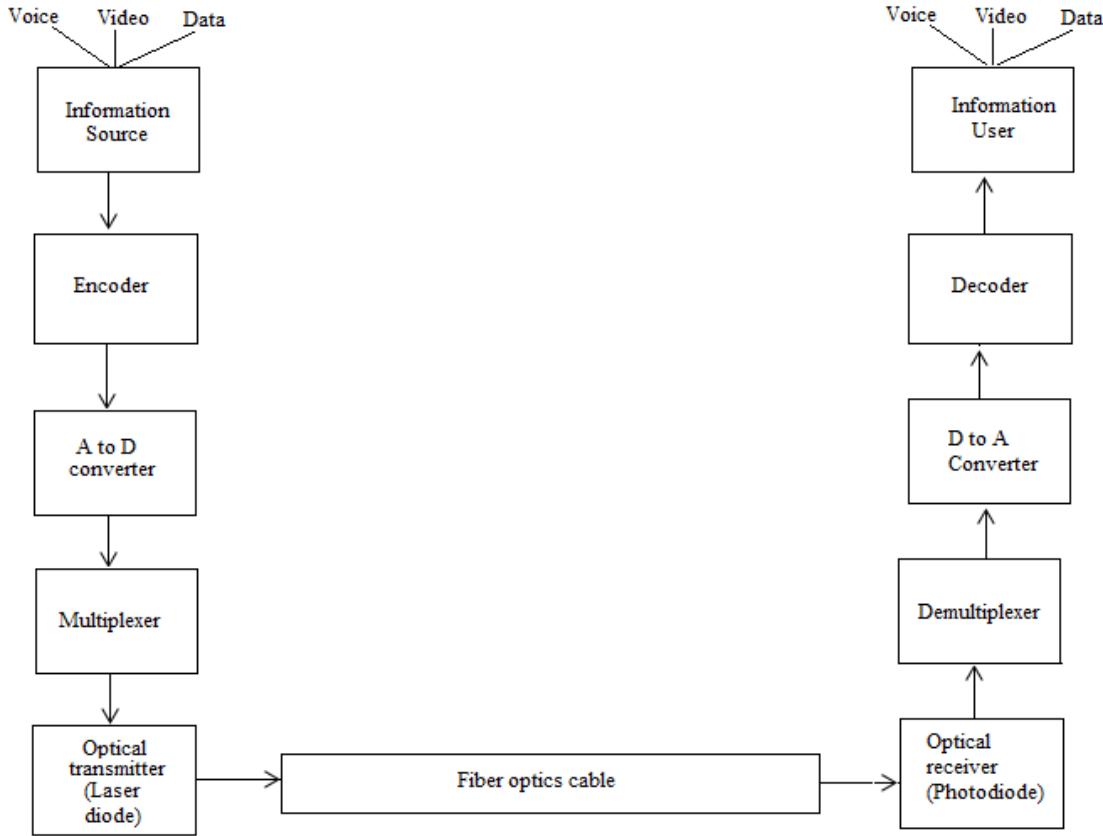


Figure (4.19) Schematic of the setup of Fiber Optics Communication System

cladding is incident on the boundary at an angle (θ) greater than critical angle ($\theta_c = \sin^{-1} \frac{\mu_2}{\mu_1}$), then it is totally internally reflected towards the core. If the optical fiber is appropriately designed then, the laser can be guided through the fiber, even if it curved.

Refer Fig (4.19), which shows block diagram of optical fiber communication network. The information such as voice, video and data can be carried through the fibers. The information is at first encoded that is converted in to electrical signal with the help of appropriate transducer (for example microphone for acoustic signals). This information is in the form of analog signal. This analog signal is converted in to digital form (sequence of bits). Various digital signals are then multiplexed (added together to form a single signal) with the help of multiplexer. The multiplexed signal in the form of sequence of '1' and '0' is then fed to optical transmitter (laser diode). In the bit sequence 1 indicates presence of the pulse, while 0 indicates it's absence. Thus '1' which indicates the presence of electrical pulse makes the laser diode ON. Thus an optical flash is emitted. '0' indicates the absence of electrical signal; therefore the laser diode remains off. Thus no optical flash is emitted. Thus the sequence of electrical bits is exactly converted in to a similar bit sequence of optical pulses. This information is then passed through the optical fibers from one place to another. Depending upon the requirement, the optical fibers may have a length of a few meters or few kilometers also. During the passage of the information the intensity

of the laser pulses decreases due to scattering and absorption. The pulses are raised to their original strength with the help of repeaters. At the receiving end, the devices work in exactly opposite manner when compared to the devices in the transmitting section. The optical pulses are converted into electrical pulses with the help of optical receiver. The optical receiver is in the form of a photodiode which gives the electrical output when the optical pulse falls on it. When the optical pulse is absent, the photodiode is off and it gives zero (output). The signals are then de-multiplexed that is separated with the help of de-multiplexer. The signals are then passed through D to A converter which converts the digital signal into analog signal. The analog signals are then converted into voice, video and data, with the help of decoder (a suitable transducer, for example, speaker for sound signals).

Fiber optic communication system has several advantages over conventional communication systems based on radio waves, microwaves and copper cables. These are discussed below

1. The capacity of the carrier for carrying the information depends upon its frequency. The frequency of light is 10^{15} Hz, while for radio waves and microwaves it is 10^6 Hz and 10^9 Hz respectively. The light offers higher bandwidth and therefore the higher number of channels.
2. The material used for optical fibers is either glass or plastic, which is less costly as compared to copper cables.
3. The laser is bright and powerful. Therefore the losses through the optical fibers are less as compared to those in copper cables. This requires less number of repeaters. In conventional communication systems, repeaters are required every 2 km while in optical network, they are required at every 100 km. Consequently the overall cost in the long run is comparatively less. The requirement of the less number of repeaters also supports the speed of the communication. The optical communication is faster than conventional communication.
4. Due to large communication carrying capacity of the light, huge information can be passed through thin optical fibers. The optical fibers thus occupy lesser space as compared to copper cables. They are also less heavy as compared to copper cables.
5. The optical fiber can withstand corrosion, changes in temperature and humidity. Optical fibers have longer life as compared to copper cables. This also makes the optical communication comparatively cheaper as compared to conventional communication systems in the long run.
6. Optical fibers have smaller size. They are flexible and yet strong. They are also light in weight.
7. The optical fiber network is free from short circuits and sparking. Thus it is safe.
8. The information carried by one fiber is not induced into another and thus cross talks are avoided. The communication is therefore more secure and private.
9. As the optical fibers are made from electrically insulating materials. The signals do not pick the electromagnetic disturbances such as those across nuclear reactors, heavy transformers, lightning and hospitals.
10. Due to the insulating properties of optical fibers, an electrical isolation in between the transmitting system and the receiving is possible.
11. As laser is absolutely coherent, the optical communication is a noiseless communication.

Fiber optic communication is used in Local Area Networks (LAN), where several computers having relatively small distance in between them are connected with optical fibers. It is also used for long distance communication. The longest optical fiber communication network is Fiber-Optic Link Around the Globe (FLAG) which is a 28,000-kilometer-long optical fiber passing through Atlantic sea. It connects the United Kingdom, Japan, and many places in between.

Optical fibers are also used as sensors for measuring the pressure, temperature, strain, magnetic field etc. Another important application of optical fibers is endoscope (used to visualize the internal organs of the patients) and fiberscopes (used to visualize the internal parts of the complex machines).



Dennis Gabor (1900 - 1979). He was a British Physicist, most notable for inventing holography, for which he later received the 1971 Nobel Prize in Physics. He studied at Technical University of Berlin. He was a Professor of Physics at Imperial college, London, where he invented Holography. After the invention of laser, holography found many applications in science and technology. In the subsequent sections we discuss the outline of holographic principles.

(4.9) HOLOGRAPHY: (Compulsory)



A promise for superior information storage

Holography is a technique of producing three dimensional images and photographs. In Greek language, holos means complete and graphien means recording. As we know, conventional photographs are two dimensional while the real objects are three dimensional. Thus, in conventional photography, information about third dimension is lost. The photographs represent incomplete information about the object. This is because, in conventional photography, the information of only amplitude of object beam is recorded. In 1948, Dennis Gabor from Imperial College, London outlined the idea behind holography. Unlike conventional photography, in holography, the information about the amplitude as well as phase of the object beam is recorded. The holograms thus represent the object completely. Holograms are as true as object itself. Holography is based on the concept of interference. As interference requires coherence, the holographic technique was realized in practice only after the invention of the laser in 1960. In this section, we discuss the basic physics behind the holography and it's applications.

Holography is a two-step process. The first step is construction (recording) and the second is reconstruction. In both these steps, lens is not required. Hence holography is called as lens-less photography.

Hologram construction (recording): Refer Fig (4.20). A laser beam is made to fall on a mirror. Another part of the same laser beam is incident on the object. The mirror reflects the beam towards a photographic plate. This beam is called as *reference beam*. The beam scattered from the object is called as *object beam*. The object beam contains information about amplitude and phase of the object. The object beam has phase variation according to the structure of the object.

The object beam is also directed towards the photographic plate. Both object beam and reference beam are coherent as they are derived from the same laser beam. On, photographic plate there occurs an interference of the reference beam and object beam. As interference is a phase dependent phenomenon, the interference of the object beam with the reference beam records the phase variation of the object. The interference pattern on the photographic plate is a complex pattern of the maxima and minima. It serves as a complex diffraction grating. This complex interference patterns does not look like the object. It is not the direct image of the object. Such photographic plate is also called as hologram. For visualizing the object in its true form, the hologram needs to be illuminated by a beam exactly identical to the reference beam.

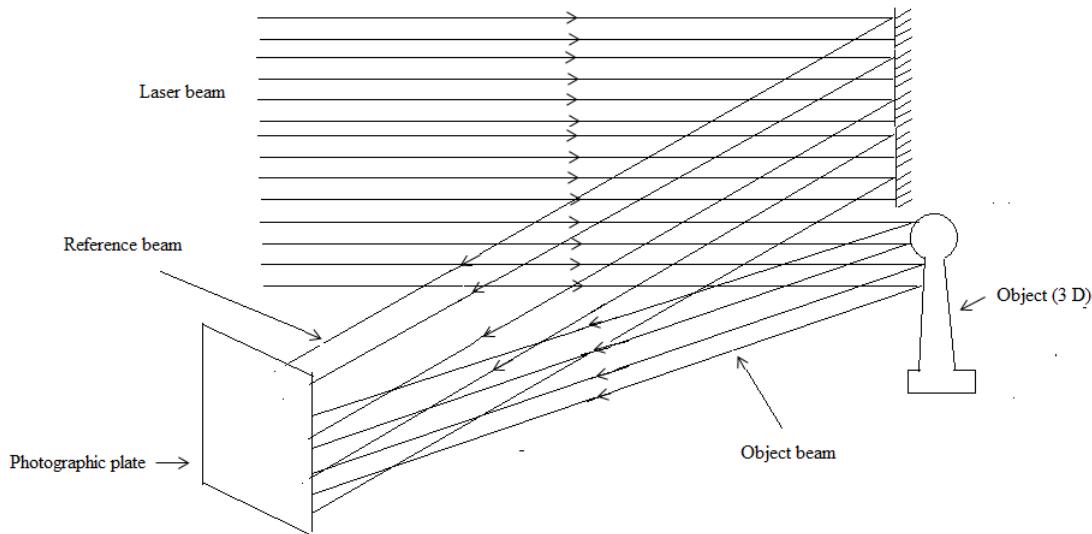


Figure (4.20): Hologram construction

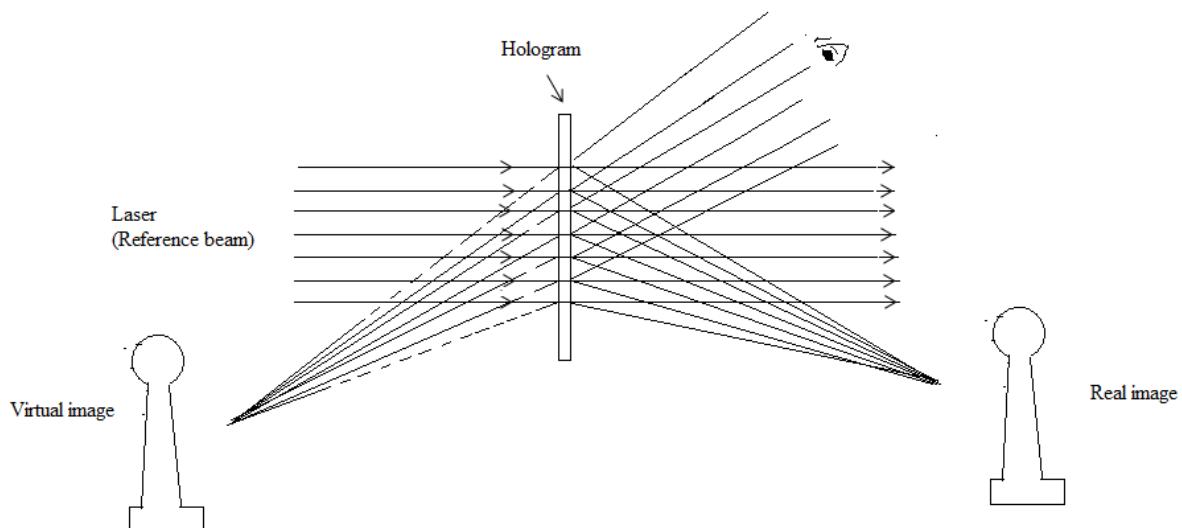


Figure: (4.21) Hologram reconstruction

Hologram reconstruction: As discussed earlier, if viewed in ordinary light, the hologram does not appear like an object but like a complex diffraction grating. In order to have a 3D perspective of the object, the hologram is to be illuminated by the same laser which was used as a reference beam, during construction. When illuminated by such laser, the laser is diffracted from the complex diffraction grating. As shown in the Fig (4.21), this produces two images, one virtual and another real. The virtual image appears at the same position, as it was occupied by the object (or on the same side of the reference beam, opposite to the observer. It is also called as true image. The real image is produced in front of the hologram (opposite side of the reference beam, or at the same position as that of observer). The real image can be photographed while virtual image is only for viewing.

One of the key features of holographic image is parallax. We know that when any real object is observed at different angles, it appears differently that is, the perspective . This is parallax. This does not happen for ordinary photographs. Even if the photograph is viewed at the different angles it does not appear differently. There is no parallax. On the contrary, the holographic 3 D image, when observed at different angles, appears differently. Holograms just resemble like the original object. Another interesting and useful feature of the hologram is that, even if it is broken into pieces, each piece contains the same amount of information, as possessed by the original hologram. The entire image can be reconstructed from the piece also. Holographic sensitive to vibrations, it is to be recorded only in the vibration free environments.

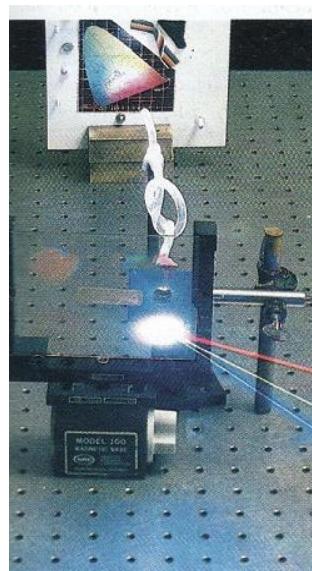
Difference between the ordinary photography and the holography:

1. In photographs there is no parallax, but in holography it is.
2. For recording the photograph, ordinary light is sufficient, while for holography (recording and reconstruction) laser is necessary
3. For holography, there is a necessity of interference between object beam and the reference beam, but this is not necessary in ordinary photography.
4. It is not possible to store different photographs on the same photographic plate, while different images or different data can be stored on the same hologram at different angles. This increases the data storage capacity by several orders of magnitude.
5. If a photograph is broken in to pieces, the information is lost, while as discussed earlier, even if hologram is broken in to pieces, the entire information can be retrieved from a single broken piece also.
6. The information stored on the hologram can be retrieved only if it is illuminated by the same laser which was used as a reference beam, thus it is possible to store the information in a secret manner. The ordinary photographs have nothing secret.

Applications of Holography:

1. Holograms are used to store the information or data. The storage capacity of holograms is considerably higher than the conventional techniques. In ordinary CD or DVD, the information is stored at one angle only, while in the holographic data storage, the information is stored at each possible angle. This enhances the data storage capacity and called as volume data storage.

2. Holography is also used in the holographic trademarks. For viewing such holograms three dimensionally, ordinary white light can also be used. This is called as rainbow holography.
3. Holography is also used for studying the air pollution.
4. Acoustic holography combines the principles of the holography with those of ultrasonography. 3 D images of the internal organs or the embryo can be obtained. The range of sound waves in the dense objects is more as compared to light. Moreover, the objects which are opaque to light can be visualized with the help of acoustic holography
5. Holographic microscopy is used to visualize blood cells, cancer affected tissues, microorganisms etc. Smaller areas of the object can be inspected with great details.
6. Holographic pattern recognition is used identify fingerprints etc.
7. Holographic trademarks are used to avoid counterfeiting. The secrecy, security and product identification is possible
8. Holographic interferometry is used to visualize the minute deformations in the object. This is useful in vibrational analysis, stress-strain analysis, and structural analysis.
9. Holographic Optical Elements (HOE) are the holograms which can be used as lenses, filters, beam splitters, diffraction gratings and the other optical elements. Same hologram can be used to function as different optical elements
10. Holography has significant applications in medical field also. Holographic endoscopes provide 3D perspective of the internal organs of the human body. Holography is also used in ophthalmology, orthopedics etc.



A holographic record diffraction grating can be used to combine primary colors from independent lasers into one “white light” for making full-color holograms

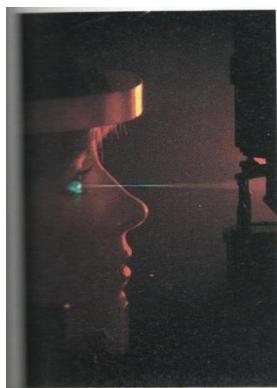
Other applications of laser in information technology:

Lasers printers are known for their speed, quality and graphics printing. Lasers are used for reading and writing a CD. The information is written and read in the digital format that is ‘1’s and “0”. Semiconductor lasers are employed for this purpose.

Applications of laser in medical field:

The focusability of laser within an extremely tiny region which increase it's energy density makes it applicable in many areas of medical field. Such finely focused laser beam is used as cutting and cauterizing element. Varieties of lasers are used in medical field. These include, CO₂ lasers, semiconductor diode lasers, dye lasers, excimer lasers, gas lasers, free electron lasers etc.

1. **Ophthalmology:** Lasers can be used for eye surgeries. The detached retinas can be welded with the help of laser beam. Lasers are also used in cataract surgeries and welding corona etc.
2. The plaque results in to narrowing of the artery. This leads to into blood pressure and heart attacks. Such plaques, the obstructions in the veins, blood clots and can be removed with the help of laser beam. The intense heat results in vaporization of the plaque or obstruction
3. Plastic surgeries
4. Laser is also used for bloodless cancer surgeries. As the laser can be focused into a fine spot, only diseased tissue is removed. The surgeries are bloodless, as the heat generated due to laser results into on the spot welding of the veins. This is called as cauterizing
5. Lasers are also used for painless dental surgeries. The local anesthesia can be avoided



Laser beam is sent into the eye of a diabetes patient to seal blood vessels in the retina



Eye surgery by using laser guided through optical fibers

Industrial applications of laser:

The focussibility of lasers allows concentrating high power within the tiny region. High power lasers such as CO₂, and ND-Yag lasers are used for industrial applications.

Welding: The intense heat generated due to a well-focused laser is useful for welding purposes. The work pieces, such as the two metal plates are held in contact. The well-focused laser beam is moved across the contact line. The metals are heated up to melting point. They stick together when cooled. The welding is carried out in the presence of gases such as argon or helium. This prevents the oxidation.

Laser welding is advantageous over conventional welding in two respects. One is local heating. The welding is done on the spot and hence very small area of the metals near the welding region is affected due to heat. Secondly as this procedure is contactless, no impurities are introduced. The laser welding is employed in automobile, aircraft and shipping industry.

Cutting: Work piece which is to be cut is exposed to well-focused laser beam. Oxygen is also blown on the spot. The region exposed to laser is burnt. The presence of oxygen supports the burning process. The oxygen also blows away the material. The burning process is enhanced due to oxygen blowing away the vaporized part.

The advantages of laser cutting over the conventional methods are speed and precision. In garment industry, several layers can be cut on the spot without frayed edges. The metals such as brass, aluminum as well as copper can be cut. Similarly, the components such as

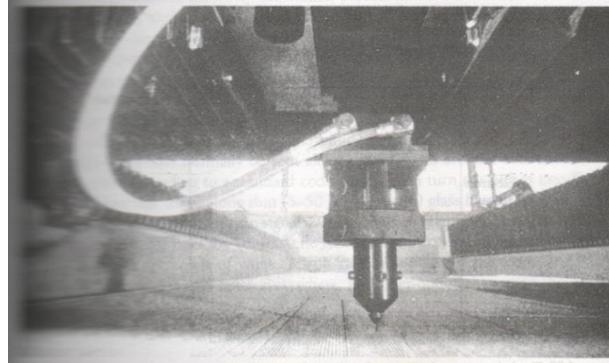
Drilling: The basic principle of laser drilling is vaporization. Very small holes having dimensions of the order of a few microns can be drilled without affecting the nearby area. The vaporized part is blown away with the help of gas jet. Pulsed lasers are used for this purpose. The procedure is free from vibration. The advantages of laser drilling are speed, accuracy and high aspect ratio (the ratio of the length/depth of the hole to its diameter). The drilling can be done in the delicate components such as PCB as well as the hard materials such as diamond. This process is also called as *micromachining*.

Heat treatment: In automobile industry the metals and other materials and the work pieces such as cylinder blocks, gears and camshafts can be hardened with the help of heat treatment. This process is also called as *laser processing*.

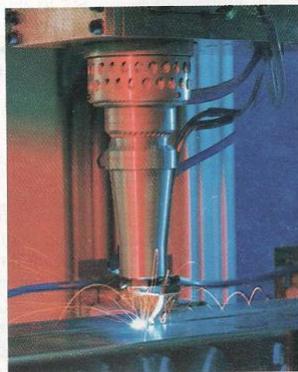
Other applications of laser:

1. Lasers are used for surveying and ranging. The intensity and directionality is useful in measuring the large distances. The long distances such as earth to moon distance have been measured with the help of laser with accuracy. The ranging with the help of laser is also called as *LIDAR*. LIDARS are also used for measuring the velocity, direction and size of the objects in the sky.
2. The laser can be used to inspect the pollution level in the atmosphere. The pollution may be due to suspended particles such as dust and smoke and gases such as carbon monoxide, sulfur dioxide etc. The concentration of the pollutants is measured as a function of the distance. A LIDAR sends a beam towards the atmosphere. The laser beam scattered from the atmosphere is analyzed. Unlike the conventional methods, LIDAR provides a real time analysis. The detection of gases depends upon the principle of Raman Effect. When the laser is scattered due to a gas, its wavelength changes. This change is called as Raman shift and it is specific for a particular gas. Ozone concentrations can be measured by this method.

3. **Barcode scanning:** He-Ne lasers or semiconductor laser is used for this purpose. The laser which scans the code is reflected back, detected by the photodetector and then fed to the computer.



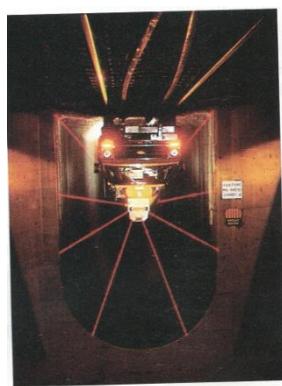
Laser being used for cutting the fabric in the garment industry



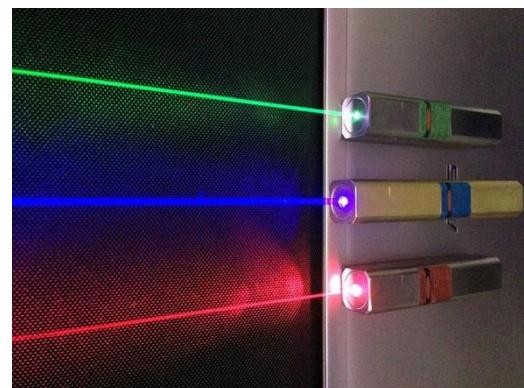
A high power laser is used for welding



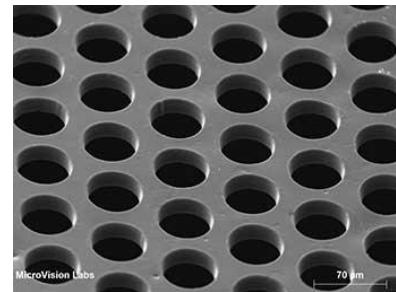
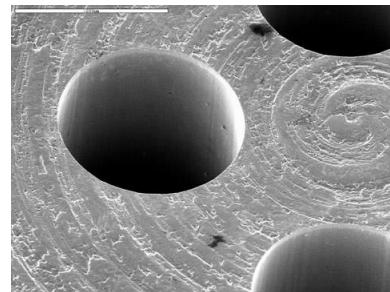
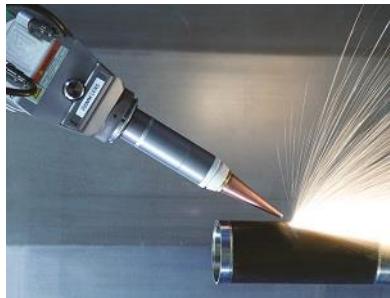
Carbon dioxide laser takes only two minutes to cut out a steel saw blade



A Helium Neon laser mounted on a truck is used to precisely measure the clearance in a tunnel



Semiconductor lasers with different colors



Laser being used for drilling the holes. Look at the precision

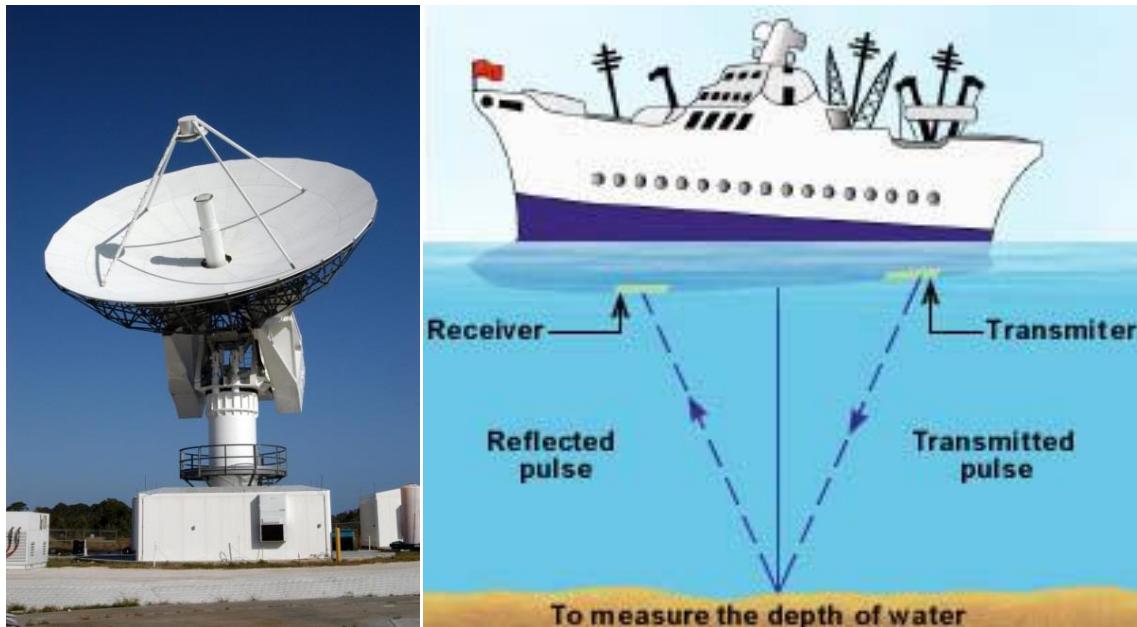
REFERENCE BOOKS

1. Principles of lasers – Orazio Svelto – Plenum Publishing Corporation, New York, 1982
2. Lasers fundamentals – W.T. Silfvast.
3. Laser and non-linear optics – B.B. Laud (2nd Edition).
4. Lasers – A.G. Sigman, Oxford University Press 1986.
5. Principles of laser and their applications – Callen O’Shea, Rhodes.
6. An introduction to Laser theory and application – M.N. Avdhanulu – S.Chand Publications.

WORLD WIDE WEB

1. <https://www.laserworld.com/>
2. www.lasertech.com/
3. Indian laser association (www.ila.org.in/)

Sound Engineering II: Ultrasonics



The image on the left side shows a RADAR, while the figure on the right indicates the schematic of a SONAR. RADARs are based on radio waves and they are used to detect and range the objects in the sky such as aeroplanes, helicopters etc. SONARs are based on ultrasonic waves and are used to navigate and range the objects in sea such as submarines, shoal of fishes etc. The RADAR and SONARs can not interchange the role of each other. Why?

The answer is in this chapter

Index

3.1 ULTRASONICS; AN INTRODUCTION:

What makes the soundless sound vastly applicable in technology?

3.2 PRODUCTION OF ULTRASONIC WAVES: MAGNETOSTRICTION OSCILLATOR

How to convert magnetic energy into mechanical vibrations

3.3 PRODUCTION OF ULTRASONIC WAVES: PIEZOELECTRIC OSCILLATOR

How to convert electrical energy into mechanical vibrations

3.4 METHODS FOR DETECTION OF ULTRASONIC WAVES

They are based on conversion of ultrasonic energy into other forms

3.5 APPLICATIONS OF ULTRASONIC WAVES

They range from SONARs to ultrasonography

3.1 ULTRASONICS; AN INTRODUCTION

What makes the soundless sound vastly applicable in technology?

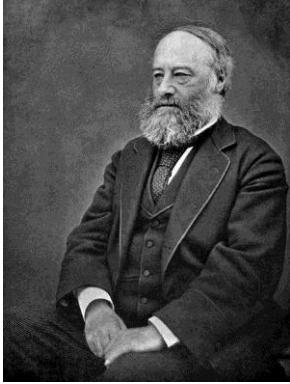
Sound waves are longitudinal vibrations. They propagate in the medium in the form of compressions and rarefactions. The sound waves in the frequency range from approximately 20 Hz to 20 KHz are audible to human beings. Hence this range is called as audible range. The sound waves having frequencies less than 20 Hz are known as infrasonic waves while the sound waves having frequencies greater than 20 kHz are known as ultrasonic waves. Both infrasonic waves and ultrasonic waves are not audible to human beings. However the ultrasonic waves having frequencies greater than 20 kHz are audible to dogs, bats, dolphins and certain insects. Bats emit and detect ultrasonic waves to find their ways even in the dark. Dolphins use ultrasonic waves to detect their prey and obstacles in their paths. Ultrasonic waves are especially useful to dolphins below the sea water where the visibility is very less.

Following properties of ultrasonic waves makes them vastly applicable in engineering, technology, medicine and science.

- i. Ultrasonic waves have very less wavelength (sometimes fraction of a mm). Therefore they exhibit negligible diffraction effects. Thus they can travel over large distances as a narrow beam. This property is particularly useful in SONARS.
- ii. Ultrasonic waves have high frequency. Such high frequency vibrations are useful in flying away the mosquitoes, killing insects, removing fog, and removing dirt on the cloths.
- iii. Ultrasonic waves are very energetic and powerful. The power density of ultrasonic waves can reach up to 10 kW/m^2 .
- iv. Ultrasonic waves are reflected when medium changes. This property is called as echo. This property is particularly useful in ultrasonography and SONARS (echo sounding).
- v. Ultrasonic waves produce cavitation effects when passed through liquids. This property is useful in cleaning the cloths, mixing the immiscible liquids etc.
- vi. Ultrasonic waves produce heating effect in the liquids through which they pass. This property is particularly useful in sonochemistry and in ultrasonic massaging massage massaging

Ultrasonic waves are widely useful in medical diagnostics and therapy, marine applications, nondestructive testing of finished products and so on.

Unlike audible sound waves, ultrasonic waves cannot be generated by vocal chords and neither they can be detected by human ears.. As they have high frequency, it requires special technology for their generation and detection. This chapter deals with the methods of generating and detecting ultrasonic waves and discusses their few important applications such as nondestructive testing, cavitation, measurement of gauge, medicine etc.



James Perscott Joule (1818-1889): He was an English Physicist. He made several contributions in Physics such as establishing a correlation between heat and mechanical work (mechanical equivalent of heat), law of conservation of energy, first law of thermodynamics etc. He, along with Kelvin, also developed Kelvin scale for the temperature measurements. He also derived a relationship between the current through the resistor and heat dissipated, which is called as Joule's first law ($E = I^2R$). He declared magnetostriction effect in 1842. Joule was a student of John Dalton and contemporary of Kelvin, Helmholtz, Stokes, Faraday and Thompson. Several honors and awards were bestowed on him. SI unit of energy is named after James Joule.

3.2 PRODUCTION OF ULTRASONIC WAVES: MAGNETOSTRICTION OSCILLATOR

How to covert magnetic energy in to mechanical vibrations

Magnetostriction effect: Magnetostriction effect was discovered by Joule in 1847. According to this effect, if a rod of ferromagnetic substance such as nickel or iron is placed in magnetic field parallel to its axis, its length changes. This change in length does not depend upon the direction of the magnetic field, but depends upon its magnitude and nature of material. The magnetostriction effect occurs more strongly in nickel than in iron. If a coil is wrapped around the rod and if a current is passed through it then a magnetic field is set up parallel to its axis. If the rod of nickel is kept in alternating magnetic field, then in each half of the cycle, as the magnitude of the magnetic field changes, the rod expands and contracts and thus vibrates mechanically. Thus the frequency of mechanical vibrations is twice the frequency of alternating magnetic field. The magnitude of the vibrations is generally very small, however, if the frequency of alternating magnetic field matches with the natural frequency of vibration of the rod then the resonance occurs and the amplitude of vibrations becomes considerably larger. The natural frequency of vibrations of the rod is given by

$$f = \frac{P}{2l} \sqrt{\frac{Y}{\rho}} \quad \dots(3.1)$$

Where P is an integer with values such as 1,2,3 ..., l is the length of the rod, Y is the Young's modulus and ρ is the density of the rod.

Example (3.1)

Calculate the natural frequency of a 50 mm length of pure iron rod. Given the density of pure iron is $7.25 \times 10^3 \text{ kg/m}^3$ and its Young's modulus is $115 \times 10^9 \text{ N/m}^2$. Can this rod be used to produce ultrasonic waves?

Solution:

The natural frequency of vibration is given by

$$f = \frac{P}{2l} \sqrt{\frac{Y}{\rho}}$$

Substituting P = 1 for fundamental frequency and the values of Y, ρ and l

$$f = \frac{1}{2 \times 50 \times 10^{-3}} \times \sqrt{\frac{115 \times 10^9}{7.25 \times 10^3}} = 39.83 \text{ kHz}$$

The frequency of vibrations is 39.83 kHz. This is greater than 20 kHz. Thus ultrasonic waves can be produced.

Example (3.2)

Calculate the length of the iron rod required for producing ultrasonic waves of frequency 25 kHz. Given the Young's modulus of iron = $11.5 \times 10^{10} \text{ N/m}^2$ and density of iron = 7250 kg/m^3 .

$$f = \frac{P}{2l} \sqrt{\frac{Y}{\rho}}$$

Considering fundamental frequency (P=1)

$$l = \frac{P}{2f} \sqrt{\frac{Y}{\rho}}$$

$$l = \frac{1}{2 \times 25 \times 10^3} \sqrt{\frac{11.5 \times 10^{10}}{7250}}$$

$$l = 0.0797 \text{ m}$$

$$l = 7.97 \text{ cm}$$



George Washington Pierce (1872–1956): He was an American Physicist. He worked as a Professor of Physics in Harvard University and is mainly known as an inventor. He wrote several books, published several research papers and received 58 patents. His main contribution is a crystal oscillator, which revolutionized the communication technology. He received his Ph.D. in 1900 and then was invited as a lecturer in Harvard University. His key contribution in wireless telegraphy was the concept of resonance and modulation. In 1914 he was assigned directorship of Cruft Physics Laboratory at Harvard and then in 1917 he became full-fledged professor in Harvard itself. He also wrote two books namely *Principles of Wireless Telegraph and Electric Oscillations and Electric Waves* which played a major role in contemporary communication technology. The magnetostriction oscillator, which is discussed below, has been designed by G. W Pierce.

Magnetostriction oscillator: (Optional)

The magnetostriction oscillator was at first designed by G.W. Pierce. The magnetostriction oscillator is shown in Fig 3-1.

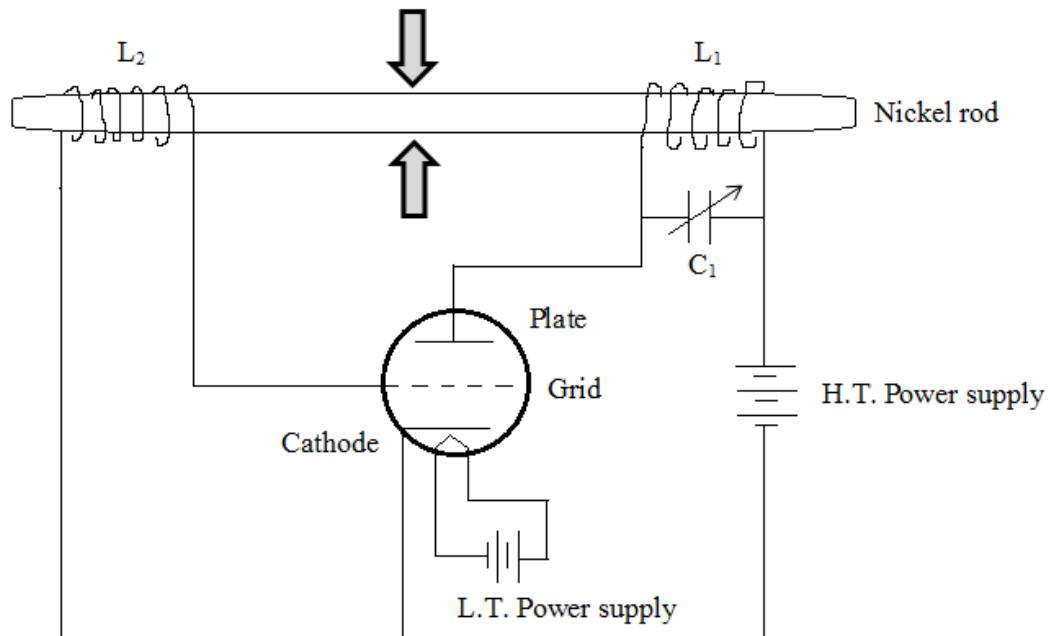


Figure 3.1: Magnetostriction Oscillator

The nickel rod is clamped at the centre. Inductance coils L_1 and L_2 are wrapped around the rod at two ends. Coil L_1 and capacitor C_1 constitute tank circuit. The capacitor C_1 is variable. By adjusting the capacitance C_1 , the frequency of the oscillation can be adjusted according to the following Eqn

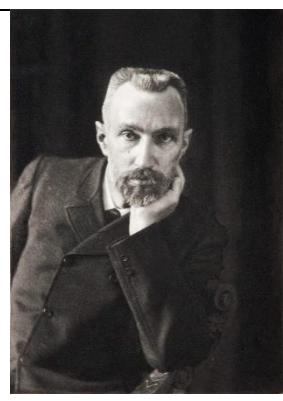
$$f_{LC} = \frac{1}{2\pi\sqrt{LC}}$$

The coil L_2 is a feedback coil and is included in the cathode-grid circuit of the valve. The coil L_1 is included in the grid-plate circuit of the valve. Coils L_1 and L_2 are inductively coupled. The oscillations of required frequency are generated by L_1C_1 tank circuit. Due to this, the current in the coil changes alternatively. Due to change in current the magnetic field in L_1 changes alternatively. Due to change in magnetic field, the rod contracts and expands alternatively according to magnetostriction effect. Due to alternate compression and expansion of the rod, the magnetic flux in the rod changes alternatively. This creates $\frac{d\phi}{dt}$ in the coil L_2 . Due to this, an emf is induced in L_2 according to Lenze's law. This emf is fed back to the cathode-grid circuit of the triode valve, which is then amplified. The amplified emf across grid-plate circuit is fed back to the L_1C_1 tank circuit and thus the oscillations are maintained (Note that the L_1C_1 circuit itself is an oscillator, however, it suffers damping due to resistance of L_1 and C_1 . The triode amplifier and the feedback mechanism is used to maintain the oscillations with constant amplitude). When the frequency of the oscillator matches with the natural frequency of the vibration of the Nickel rod, resonance takes place and the rod vibrates with maximum amplitude.

Advantages of magnetostriction oscillator are

- i. Magnetosctriction oscillators are inexpensive
- ii. Large output power can be produced

The disadvantage is that the frequency limit is 300 kHz. In order to produce ultrasonic waves in MHz to GHz range, a piezoelectric oscillator is required



Pierre Curie (1859-1906): He was a French physicist, a pioneer in many areas of Physics such as crystallography, magnetism, piezoelectricity and radioactivity. In 1903 he received the Nobel Prize in Physics with his wife, Marie Curie, "in recognition of the extraordinary services they rendered by their joint researches on the radiation phenomena discovered by Professor Henri Becquerel". He studied paramagnetism, diamagnetism and ferromagnetism extensively as a part of their doctoral thesis, where he established Curie law (effect of temperature on paramagnetism) and Curie point (where the ferromagnetism is destroyed at critical temperature). Mary Curie and Pierre Curie were the pioneers in the field of radioactivity. Curie along with his student discovered beta, gamma and alpha rays. His name has been given to the unit of radioactivity. His daughter Irèneand Joliot-Curie won a Nobel Prize in Chemistry in 1935. The entire Curie family worked as Physicists. In 1880, Pierre and his brother Jacques discovered direct as well as indirect piezoelectric effect, which is discussed in next paragraphs

3.3 PRODUCTION OF ULTRASONIC WAVES: PIEZOELECTRIC OSCILLATOR (Compulsory)

How to convert electrical energy into mechanical vibrations

Piezoelectric effect: Piezoelectric effect was discovered by French Physicists Pierre Curie and Paul-Jean Curie in 1880. Piezoelectricity means pressure electricity. This means that, with the help of certain crystals, pressure can be converted into electricity or conversely electricity can be converted into pressure. Refer Fig 3-2. If one pair of opposite faces of quartz crystal is applied

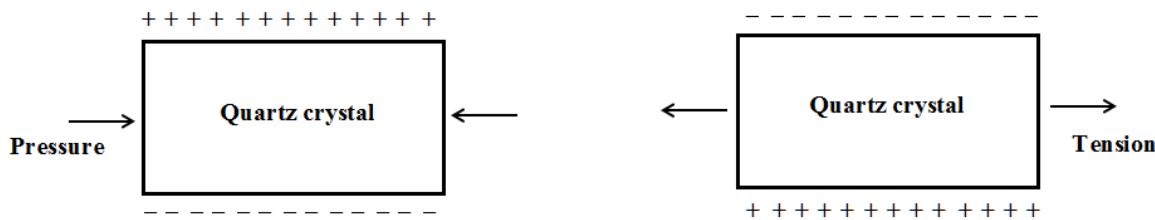


Figure 3.2: Piezoelectric effect in Quartz crystal

with pressure then equal and opposite charges are developed across the face perpendicular to which pressure is applied. If tension is applied instead of pressure then the sign of the charges is altered. The magnitude of the electricity produced is proportional to the pressure or tension applied. The conversion of pressure (or tension) into electricity is called direct piezoelectric effect. The converse of piezoelectric effect also occurs. This means that if DC electricity is applied to a pair of opposite faces of the crystal then the dimension of the crystal changes. If the sign of the charges is altered then the dimension changes in opposite manner. This is known as inverse piezoelectric effect. Piezoelectric effect is exhibited by asymmetric crystals such as quartz, tourmaline, rochelle salt, ammonium phosphate, PZT (lead zirconate titanate) etc. If the crystal is placed in alternating electric field then it contracts and expands and thus if the frequency of the AC voltage is suitably adjusted then the vibrations of crystal produce ultrasonic waves in the medium. If the frequency of the applied AC voltage is matched with one of the natural modes of vibrations of the crystal, then the resonance occurs and the crystal vibrates with maximum amplitude. The frequency of the thickness vibrations is given by

$$f = \frac{P}{2t} \sqrt{\frac{Y}{\rho}}$$

The frequency of length vibrations is given by

$$f = \frac{P}{2l} \sqrt{\frac{Y}{\rho}}$$

Where P is an integer having values 1,2,3 for fundamental, first overtone, second overtone respectively, t is the thickness and l is the length of the crystal. Y is the Young's modulus and ρ is the density of the crystal.

Geometry of Quartz crystal: Amongst crystals which exhibit piezoelectric effect, quartz is generally preferred for the piezoelectric oscillators. Naturally occurring quartz is a chemically as well as physically stable material. It is hard. It is in the form of hexagonal prism with pyramid at both ends (refer Fig 3-3 (a)). The line joining two opposite ends is the axis of symmetry and is called as optic axis.

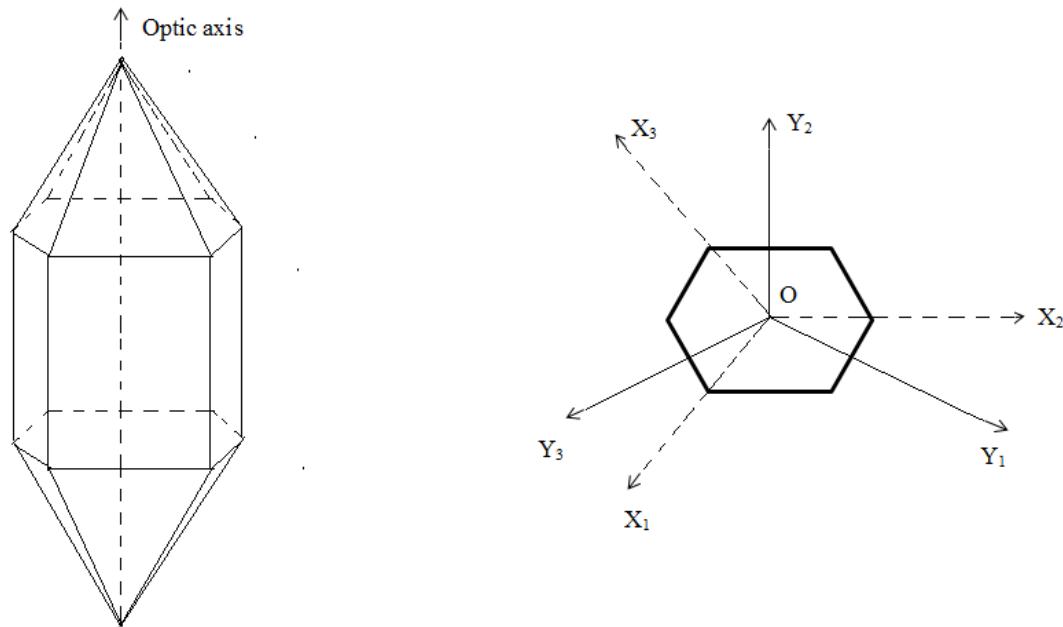


Figure 3-3(a) Geometry of Quartz crystal (b) Electric and mechanical axes of the Quartz crystal

The crosssection of the crystal is a regular hexagon (refer Fig 3-3 (b)). The three axes which pass through the opposite corners are called as X axes and they are electrical axes. The three axes which are perpendicular to opposite faces are called as Y axes and they are mechanical axes. Three electrical axes and the three mechanical axes are perpendicular to each other. The plate of the crystal cut with its faces perpendicular to X axes is called as X -cut plate, while the plate of the crystal cut with its faces perpendicular to Y axis is called as Y -cut plate. When the voltage is applied in the direction of electrical axes the dimensions of the crystal change in the direction of mechanical axes. If a stress (or strain) is applied in the direction of mechanical axes then voltage is developed in the direction of electric axes.



Paul Langevin (1872-1946): He was a French Physicist. He studied at Cambridge University under Sir. J.J. Thomson, a discoverer of electron and a Nobel laureate. In 1917, he designed a piezoelectric oscillator and consequently the first SONAR for both of which he was assigned two US patents. He was a doctoral student of Pierre Curie. Langevin is also known as a proposer of the famous Twin Paradox in the special theory of relativity. He also developed Langevin dynamics and the Langevin equation. In the paragraphs below, we discuss a piezoelectric oscillator, which was first designed by Langevin.

Piezoelectric oscillator: **Compulsory**

The first piezoelectric oscillator was constructed by Langevin in 1917. Consider Fig 3-4. The tank circuit L_1C_1 and the triode amplifier constitute a high frequency oscillator, which generates a high frequency voltage of frequency given by following relation

$$f_{LC} = \frac{1}{2\pi\sqrt{LC}} \dots (3.2)$$

Thus the frequency of the applied ac voltage can be adjusted by adjusting the capacitance of the variable capacitance C_1 . The L_1C_1 tank circuit is connected to a capacitor having metal plates A and B. A thin slice of properly cut piezoelectric crystal is kept in the plates A and B as a dielectric. The triode amplifier is used for maintaining the oscillations. The triode contains three

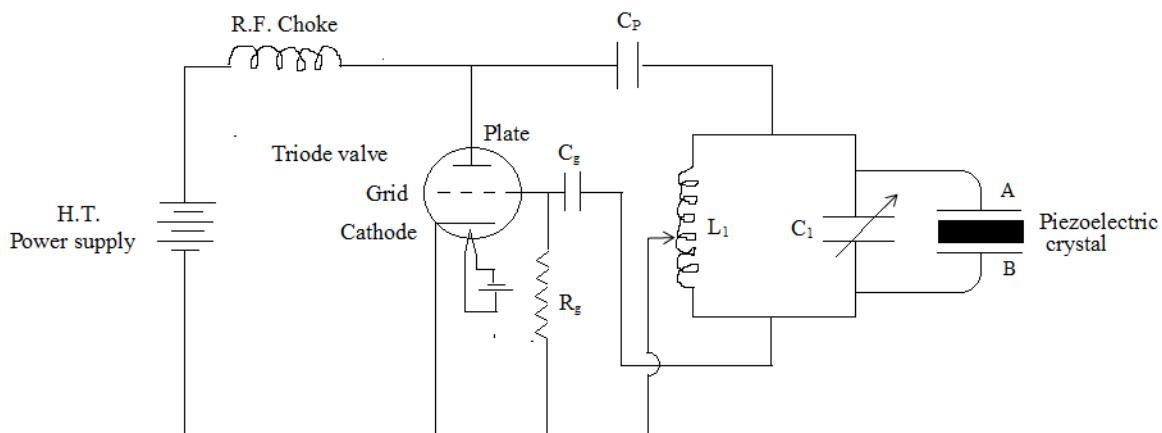


Figure 3.4 Piezoelectric oscillator

elements namely plate (P), cathode (C) and grid (G). The capacitor C_g and resistor R_g are used to bias the grid. When the frequency of the electronic oscillator matches with the natural frequency of vibration of the quartz crystal, resonance occurs and crystal vibrates with maximum amplitude. In the circuit, the RF choke is used to prevent RF frequency coming from plate circuit in

reaching the H.T. supply. The capacitor C_P is used to block DC voltage of the HT supply in reaching the L_1C_1 tank circuit, while at the same time it allows AC voltage from triode valve to reach the L_1C_1 circuit. The piezoelectric generator is used to generate the ultrasonic waves of frequency around 15×10^7 Hz.

Example (3.3)

Calculate the natural frequency of the quartz plate having thickness 5.5 mm. Given, Young's modulus of quartz is 8.0×10^{10} N/m² and density is 2.65×10^3 kg/m³.

Solution:

We have

$$f = \frac{P}{2t} \sqrt{\frac{Y}{\rho}}$$

Taking n=1 and substituting the data

$$f = \frac{1}{2 \times 5.5 \times 10^{-3}} \sqrt{\frac{8 \times 10^{10}}{2.65 \times 10^3}}$$

$$f = 49.95 \text{ kHz}$$

Example (3.4)

Find the thickness of the quartz plate required to produce ultrasonic waves of frequency i) 3 MHz and ii) 350 kHz. (Density of quartz = 2650 kg/m³, Young's modulus = 8×10^{10} N/m²).

Solution:

$$f = \frac{P}{2t} \sqrt{\frac{Y}{\rho}}$$

Considering fundamental frequency, we take P =1, thus

$$t = \frac{P}{2f} \sqrt{\frac{Y}{\rho}}$$

For f = 3 MHz

$$t = \frac{1}{2 \times 3 \times 10^6} \sqrt{\frac{8 \times 10^{10}}{2650}}$$

$$t = 9.16 \times 10^{-4} \text{ m}$$

$$t = 0.916 \text{ mm}$$

For $f = 350 \text{ kHz}$

$$t = \frac{1}{2 \times 350 \times 10^3} \sqrt{\frac{8 \times 10^{10}}{2650}}$$

$$t = 7.85 \times 10^{-3} \text{ m}$$

$$t = 7.85 \text{ mm}$$

3.4 METHODS FOR DETECTION OF ULTRASONIC WAVES

(Optional except Piezoelectric method)



They are based on conversion of ultrasonic energy in to other forms

a. Piezoelectric crystal method: The production of ultrasonic waves is based on inverse piezoelectric effect, while their detection is based on direct piezoelectric effect. In this method, when a quartz crystal is subjected to ultrasonic waves, it mechanically vibrates due to their compressions and rarefactions. As a result of piezoelectric effect, alternate charges and hence a small AC voltage is developed across the perpendicular faces. This voltage is then amplified and detected with the help of detectors like voltmeter or Cathode Ray Oscilloscope.

b. Sensitive flame method: When a steady flame is exposed to ultrasonic waves, it flickers due to nodes and antinodes of ultrasonic waves. It remains stationary at antinodes and flickers at the nodes.

c. Kundt's tube method: When a glass tube with one open end is exposed to ultrasonic waves, standing waves consisting of nodes and antinodes is formed. When the tube is clamped horizontally and if a small amount of lycopodium powder is spread at the lower surfaces, the powder forms heaps at the nodes and is blown off at antinodes. The distance between the heaps corresponds to the distance between the nodes and is $\lambda/2$, where λ is the wavelength of ultrasonic waves. This method is applicable to those cases where the wavelength of ultrasonic waves is not too small. (Typically the wavelength of ultrasonic waves is in fraction of mm)

d. Thermal detector method: When a fine platinum wire is subjected to ultrasonic waves, its temperature changes and hence the resistance changes at the nodes. If this platinum wire is included as one of the resistor in a Wheatstone's bridge, then when the platinum wire is moved through ultrasonic waves, the bridge is imbalanced at the nodes and the galvanometer shows a

e. Acoustic diffraction method: Consider a glass tube filled with the liquid column. When this liquid column is exposed to ultrasonic waves, nodes and antinodes are formed. The density of and hence the refractive index of the liquid column maximizes at the nodes. Thus the liquid becomes opaque at the nodes and remains transparent at the antinodes. Thus the liquid contains alternate opacities and transparencies and thus behaves like an acoustic diffraction grating. The grating element d is given by the distance between the successive nodes or antinodes and thus is $\lambda_u/2$, where λ_u is the wavelength of ultrasonic waves. If such acoustic grating is exposed to monochromatic source of light then a typical diffraction pattern characteristic to a diffraction grating is produced. The grating equation in such case is given by

$$ds\sin\theta = m\lambda$$

Where d is the grating element given by

$$d = \frac{\lambda_u}{2}$$

m is the order of spectrum and θ is the angle of diffraction of m^{th} order and λ is the wavelength of monochromatic source of light. Thus if λ is known and θ for a given order (m) is measured, then the grating element (d) can be calculated and thus the wavelength of ultrasonic waves can be calculated by using relation $\lambda_u = 2d$. Further, if the frequency of ultrasonic waves (f) is known then velocity of the ultrasonic waves through liquid column can be calculated using the relation $v = f \times \lambda_u$.

3.4 APPLICATIONS OF ULTRASONIC WAVES (All Compulsory)

They range from SONARs to ultrasonography

As discussed earlier the properties of ultrasonic waves which make them highly applicable in technology are their low wavelength, high frequency, power and vibrations. The applications of ultrasonic waves are numerous a few of which are discussed below.

SONARS (Echo Sounding):

SONAR is an acronym of SOund NAVigation and Ranging. SONAR was designed by Paul Langevin around 1917 (*as such, the SONAR was to be used for detecting the submarines of enemy during the first world war, but reportedly, by the time SONAR was developed, the war was over!*) Ultrasonic waves have high frequency and low wavelength and hence they can travel over longer distances as a narrow and directional beam. Ultrasonic waves can travel through long distances in water. There are two kinds of SONARs, one is ECHO SONAR and another is DOPPLER SONAR. ECHO SONAR is used to find the position of the object while DOPPLER SONAR is used to find the position as well as velocity of object and to know whether the object is approaching or going away. SONAR technique is used to detect the icebergs, submarines, shoal of fishes, to signal the ships and to find the depth of the sea. The ECHO SONAR is based on the echo principle. When an ultrasonic wave encounters a change in medium, it is reflected back as an echo. In SONARS, ultrasonic pulse is transmitted. At the

instant of transmission, a pulse is recorded on cathode ray oscillaoscope (CRO). This pulse is reflected from the obstacle (ship, submarine, shoal of fishes, iceberg or seabed), when the reflected pulse is detected back, another pulse appears on the CRO. The time interval (t), between the transmitted pulse and reflected pulse is measured using CRO. If velocity of the ultarsonic waves in the seawater is known, then the distance (d) between the SONAR and the obstacle can be calculated using following relation.

$$d = \frac{v \times t}{2} \quad \dots(3.3)$$

The velocity of ultrasonic waves in sea water depends upon it's salinity and the temperature

$$v = v_o + 1.14S + 4.21t - 0.037t^2 \quad \dots(3.4)$$

Where S is the Salinity of sea water in gm/liter, t is the temperature in $^{\circ}\text{C}$ and v_o is the velocity of sound at 0°C which is 1510 m/s.

The Doppler SONARs are based on acoustic Doppler effect. According to this effect, if the sound is reflected from a moving obstacle then the frequency of the reflected sound is increased or decreased depending upon whether the obstacle is approaching or going away. The dolphins, whales and bats use DOPPLER SONAR principle to avoid an obstacle in the path or to detect their preys. DOPPLER SONARs are based on following equation

$$v = \frac{\Delta f v_s}{2f \cos \theta} \quad \dots(3.5)$$

Where v is the velocity of the obstacle, Δf is the Doppler shift, v_s is the veolcity of the ultrasonic waves in the medium, f is the frequency of the ultrasonic waves and θ is the angle subtended by the ultrasonic beam with the obstacle.

Example (3.5)

Calulate the velocity of ultarsound in sea water at 35°C , if it's salinity is 30 gm/liter and if the velocity of ultrasound in water is 1510 at 0°C .

Solution: We have

$$v = v_o + 1.14S + 4.21t - 0.037t^2$$

$$v = 1510 + 1.14 \times 30 + 4.21 \times 35 - 0.037 \times 30^2$$

$$v = 1510 + 1.14 \times 30 + 4.21 \times 35 - 0.037 \times 30^2$$

$$1510 + 34.2 + 147.35 - 33.3$$

$$v = 1658 \text{ m/s}$$

Example (3.6)

Two ships are anchored at some distance from each other. An ultrasonic signal is sent by two routes i.e. through water and air. The difference between times at which the signals reach the other ship is 3 seconds. If the velocity of sound in air and water is 348 m/s and 1392 m/s respectively, find distance between the ships.

Let d be the distance between two ships. Then

in air,

$$d = v_a \times t_a$$

and in water

$$d = v_w \times t_w$$

$$v_a \times t_a = v_w \times t_w$$

$$t_a = t_w \times \frac{v_w}{v_a}$$

$$t_a = t_w \times \frac{1392}{348}$$

$$t_a = 4t_w$$

We also have

$$t_a - t_w = 3$$

Thus

$$4t_w - t_w = 3$$

$$t_w = 1$$

$$d = 1 \times 1392$$

$$d = 1392 \text{ m}$$

$$t_a = 4t_w = 4 \text{ s}$$

$$d = 348 \times 4 = 1392 \text{ m}$$

Example (3.7)

An ultrasonic pulse of frequency 100 kHz is sent down towards the seabed. The echo is recorded after 0.8 sec. If the velocity of sound in sea water is 1500 m/s, calculate the depth of sea and wavelength of pulse.

Solution:

We have $v = \frac{2d}{t}$

$$d = \frac{v \times t}{2}$$

$$d = \frac{1500 \times 0.8}{2} = 600 \text{ m}$$

Also $v = f \times \lambda$

$$1500 = 100 \times 10^3 \times \lambda$$

$$\lambda = \frac{1500}{10^5} = 0.015 \text{ m} = 1.5 \text{ cm}$$

Thickness measurement: This application is also based on the principle of echo sounding. An ultrasonic transducer is attached to one end of the test piece whose thickness is to be measured. The other end is generally attached to a support. The ultrasonic transducer emits an ultrasonic pulse (by converting electrical pulse into ultrasonic pulse). At the instant at which the pulse is transmitted, a pulse appears on the time axis of CRO. The ultrasonic pulse travels through the test piece and then returns back from the other end due to its echo owing to change in the medium. When the ultrasonic pulse is received back, it is converted into an electrical pulse. At the instant of detection of the pulse, another pulse appears on CRO. If the time interval between the transmission and the reception of the pulse is t , if the thickness of the test specimen is d and if the velocity of the ultrasonic waves through the material of the test specimen is v , then the thickness of the specimen can be calculated by using the following equation

$$d = \frac{v \times t}{2}$$

For using the above equation, the velocity of the ultrasonic waves through the material of the test specimen should be known. It can be calculated by using a test specimen of known thickness. The thickness of the guages can also be calculated by using this technique. Such devices are called as ultrasonic guage meters

The above technique is especially applicable to those specimens whose only one side is accessible (such as reactor walls, or a railway track). The method is nondestructive in the sense that there is no need to drill a hole through the specimen.

Example (3.8)

The echo time of the ultrasonic pulse of ultrasonic pulse which is travelling with the velocity $5.9 \times 10^3 \text{ m/s}$ in mild steel is $7 \mu\text{s}$. Calculate the thickness of the mild steel.

Solution:

We have

$$v = \frac{2d}{t}$$

$$\Rightarrow d = \frac{vt}{2}$$

$$d = \frac{5.9 \times 10^3 \times 7 \times 10^{-6}}{2}$$

$$d = 0.02065 \text{ m} = 20.65 \text{ mm}$$

Example (3.9)

An ultrasonic gauge meter is used to measure the velocity of ultarsonic waves in brass. The thickness of the brass specimen is 11.24 cm. If the echo time measured by the device is 52.28 μs . What is the velocity of ultrasonic waves through brass?

Solution

$$\text{We have } v = \frac{2d}{t} = \frac{2 \times 11.24 \times 10^{-2}}{52.28 \times 10^{-6}} = 4.3 \times 10^3 \text{ m/s}$$

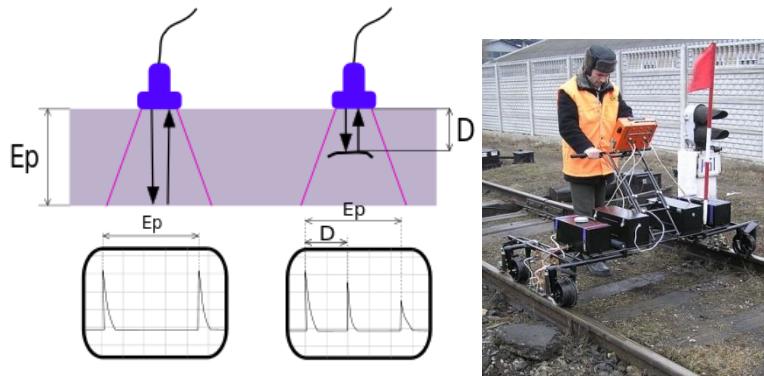
Non-Destructive Testing:

Proper functioning of several machines such as ships, railways, aeroplanes, railway tracks and industrial machines is based on the strength as well as the absence of any kind of defect in the components used in them. It is always convineient to test the defects in the components after their production and before their use in the machines. Periodic testing of the flaws at regular intervals is also necessary. Such testing saves the financial losses as well as accidents of the mechines. The defects are of various kinds suhc as cavities, discontinuity, cracks, voids, flakes, segregations etc. The testing based on ultrasonic waves is nondestructive (NDT) as it does not cause any damage in the components. Ultrasonic waves of frequency ranging from 100 kHz to 25 MHz are used for this purpose. This technique is also based on echo sounding. A piezoelectric ultrasonic transducer is suitably fixed on one end of the test piece. The ultrasonic waves/pulses are then transmitted through the piece. In absence of any defect, the ultrasonic wave/pulse returns from another end and two pulses appear on the screen of CRO. If there is a defect, an intermediate pulse appears on the screen. This indicates presence of a defect. The distance (position) the defect from the surface can be calculated by using following equation

$$v = \frac{2d}{t}$$

The shape and size of the defect can also be calculated by scanning the object from different directions. Nondestructive testing based on ultrasonic waves is advantageous than that based on gamma rays and X rays because the ultrasonic method uses a compact and less bulky apparatus. The method is relatively cheap, reliable, accurate and speedy. The objects with large sizes can be tested. Plastic components can also be tested. The defect is detectable even if it is small in size. A few disadvantages of ultrasonic testing are; it's inability to test the components of extremely

small size, necessity of smooth surfaces, and its inability to detect the defect if it is very close to the surface.



Ultrasonic Non-Destructive Testing, schematic and actual (a railway track being tested)

Ultrasonography: Ultrasonography is one of the most widely used imaging technique. It can be used to check the pregnancy or to take the images of the diseased tissues like tumor. The principle is echo sounding itself. When the ultrasonic transducer sends the high frequency ultrasonic waves through the human body, they are reflected when they encounter a structure (like tumor or foetus) having a different density (the organs in human body have different densities). The reflected ultrasonic waves are converted into digital signals which are further processed by computer to form an image by using digital processing technique. Ultrasonography technique is versatile in the sense that it can be used to find the location, size as well as displacement of the organ. Ultrasonography is also used to monitor the development of the foetus as well as to detect the defect in it, so that corrective measures can be taken. The technique is noninvasive and safe because ultrasonic waves do not cause harmful effects (such as mutations as in case of X rays). Doppler ultrasonic imaging is used to measure the flow of the blood through the heart and major arteries. It is based on the Doppler effect, according to which there is shift in frequency of the ultrasonic waves if they pass through a medium in motion.

Advantages of ultrasonic imaging:

1. Ultrasonic imaging is noninvasive. It does not use needles or injections. It also does not use ionizing radiations like X rays. It does not cause health problems and can be repeated as many times as required.
2. Ultrasonography can be safely performed without any health risk on the pregnant women.
3. Ultrasonic waves are widely available and easy-to-use. The technique is comparatively less expensive.
4. The clarity of images is better in ultrasonic imaging than in X ray imaging.



1.

Ultrasonographic image of a baby in mother's womb

Limitations of ultrasonic imaging:

2. Ultrasonic waves do not pass easily through air. Therefore they are not suitable for imaging the structures involving bowel. Due to same reason, the evaluation of stomach, small intestine and large intestine may be limited. Intestinal gas may also prevent visualization of the deeper structures such as pancreas and main artery.
3. The ultrasonic waves do not penetrate through the bones and therefore although it is used to visualize the outer surface of the bones, the internal structure can not be imaged.

Ultrasonic blood flow meter: The ultrasonic blood flow meters are used to inspect speed of blood through the veins and arteries. The principle involved in Doppler effect according to which when the ultrasonic waves pass through moving blood, their frequency is shifted. The ultrasonic transmitter and receiver is clamped on the limb. The doppler shift in the frequency Δf in the frequency of ultrasonic waves is given by the following equation

$$\Delta f = \frac{2fv\cos\theta}{v_u} \quad \dots(3.6)$$

Where Δf is the Doppler shift in the frequency, f is the frequency of incident ultrasonic waves, v is the velocity of the blood, θ is the angle subtended by the ultrasonic beam with the direction of flow of the blood, and v_u is the velocity of ultrasonic waves through the blood.

We can rearrange the above equation for v . Thus we have

$$v = \frac{\Delta f v_u}{2f \cos\theta} \quad \dots(3.7)$$

The Doppler flow meter can also be used to measure the flow of water or petrol diesel or other chemicals through the pipelines

Other medical applications of ultrasonic waves:

In medical field, ultrasonic waves can be used for diagnosis as well as therapy. The principles used are cavitation, heating effect and echo. Following are some other medical applications of ultrasonic waves

1. Curing rheumatic pains
2. Cleaning teeth
3. Painless dental cutting of teeth. The procedure does not involve any mechanical device.
4. To break kidney stones or gall stones in to small pieces so that they can easily flow out of the body
5. Ultrasonic cavitation is used to destroy tumor and cancer cells. The procedure involves no loss of blood.
6. For relaxing contracted fingers

Ultrasonic cavitation:

Cavitation means production of bubbles. The ultrasonic waves contain compressions and rarefactions. When these waves pass thorough the liquid, the compression creates a stress in the liquid and pushes the liquid apart. Thus a tiny cavity or a bubble is created. The cavity then sucks the un-dissolved gases in the liquid. The rarefaction, following after compression, creates negative pressure and allows the bubble to grow. However the growth of the bubble cannot continue indefinitely due to oppose from the hydraulic pressure of the liquid and thus it collapses. Several such microscopic bubbles are created and collapsed. Due to creation of the bubble and its collapse, a shock wave is created due to which the local pressure as well as temperature of the liquid increases to a very high value. This phenomenon is called as cavitation.

Cavitation has two negative aspects. One is that due to cavitation a dense cloud is formed near the transducer which hinders the propagation of ultrasonic waves and second is that cavitation causes creation of pits on the surface of the transducer due to which it is damaged. Despite of such negative aspects cavitation has some applications which are described below.

Ultrasonic cleaning: The creation of bubbles and their implosion sucks the dirt particles attached to the surfaces and thus helps cleaning. Ultrasonic washing machines are based on this principle. Clothes, jewelries, wrist watches, delicate machinery, metals, electronic components, integrated circuits, medical and optical instruments can be cleaned by ultrasonic cavitation method.

Ultrasonic emulsification: The immiscible liquids such as oil and water can be thoroughly mixed to form a stable emulsion by using cavitation. The mixing occurs due to implosion of the bubbles. Ultrasonic mixing due to cavitation is also used to form the stable alloys of the metals such as Iron-Lead, Aluminum-Cadmium, Zinc-Lead etc. The molten mixture of such metals is thoroughly mixed due to cavitation.

Cavitation is also used to dissolve kidney stones, for sterilization, dissolving contaminants, catalyzing chemical reactions, degassing the liquids and sonoluminescence etc.

Industrial applications of ultrasonic waves:

Ultrasonic drilling: Conventional drilling is based on the rotatory motion of the drill bits pressed against the surface to be drilled. However, ultrasonic drilling is based on the vibratory motion of the drilling bit. The piezoelectric transducer or the magnetostriction transducer is suitably connected to the drilling bit pressed on the components to be drilled. The vibrations of the piezoelectric/magnetostriction transducer are transferred to the drilling bit. Due to resonance, the vibrations are of high amplitude. The drilling bit vibrates with amplitude around 0.1 mm and frequency of 20 to 30 kHz. The vibrations of the drilling bit produce a stress in the component to be drilled and the material where the stress is produced is teared away. The advantage of ultrasonic drilling is that it does not require a large force to be applied on the drilling bit. The apparatus requires less electrical power. Ultrasonic drilling is particularly applicable for drilling in hard and brittle solids such as ceramics, glasses, precious gem stones, pearls, semiconductors and hard alloys.



Ultrasonic drilling

Ultrasonic welding: Ultrasonic welding is particularly useful in welding thin plastics (plastic toys, consumer goods, mobile accessories and automobile parts) and thin metal sheets (packaging industry). The parts to be welded are cleaned and pressed with each other and kept under a hammer connected to ultrasonic vibrator. After making ON, the hammer vibrates with ultrasonic frequency. Due to vibrations the parts are further pressed against each other as a result of which they diffuse in to each other and are welded. The welding also takes place due to friction and the heat generated in the parts being welded. The advantage of ultrasonic welding is that it does not cause the excessive stress and excessive temperature rise at the parts to be welded so that welding can be done without affecting the properties of the materials. However, ultrasonic welding requires a large power and cannot be used to weld thick metal sheets.

Ultrasonic soldering: Generally the surfaces to be soldered are covered with contaminants, grease and oxide films. These need to be removed before soldering. For this purpose, the surfaces are cleaned with active fluxes. This method is not suitable for soldering aluminum and steel. In such cases ultrasonic soldering can be used. The ultrasonic vibrations remove the oxide films and solder aluminum or steel.

Ultrasonic cleaning: Electronic components need to be quite clean at every stage of production of electronic circuits and devices. Generally organic solvents or weakly alkaline aqueous solutions containing surface active agents are used for this purpose. To clean the components more effectively, the phenomenon of cavitation can be used. The layers of contaminants are scrubbed/sucked due to the implosion of the cavitation bubbles, which creates the hydraulic shock. Bubbles penetrate under the layer, tear it off and break it down in to minute pieces. The surface active agent pulls them away into the solution. The advantage of this technique is that it can be used to clean small components. The electronic components, integrated circuits, jewelries and medical and optical instruments can be cleaned by this method.

Other applications of ultrasonic waves:

1. Ultrasonic waves can be used to accelerate the rate of chemical reactions
2. They can be used to measure the viscosity of the liquids
3. They are used for washing the clothes in textile industry
4. Dispersing fog at airports
5. Removing soot from chimneys
6. Ultrasonic coagulation is used for coagulating the loose particles of dust, mist or smoke
7. For homogenizing fat with the milk
8. Pasteurizing the milk
9. For speeding up the chemical reactions (sonochemistry)
10. It has been found that seeds subjected to ultrasound may germinate more rapidly and produce higher yields
11. For measuring the fat layers in the milk
12. For measuring molecular weight of polymers
13. To break up the molecular weight of the polymers
14. For measuring bulk modulus of the liquids by using following equation

$$v = \sqrt{\frac{K}{\rho}}$$

Where v is the velocity of ultrasonic waves through the liquid, K and ρ are the bulk modulus and density of liquid

15. For measuring adiabatic compressibility of the liquid by using following equation

$$\beta = \frac{1}{\rho v^2}$$

Where β, ρ are the adiabatic compressibility and density of liquid and v is the velocity of ultrasonic waves through liquid

16. They destroy bacteria and thus are used to sterilize water and milk
17. The animals like frog and fish can be maimed or killed using ultrasonic waves. This is because ultrasonic waves can destroy the blood vessels.
18. For agglomeration
19. For catalysis

20. For crystallization
21. For flying away insects such as mosquitoes
22. For decomposing water into H and OH

EXERCISES

Questions

1. Which properties of ultrasonic waves make them highly applicable in technology?
2. Why the applications exhibited by ultrasonic waves are not possible using audible sound?
3. What is the exact use of ultrasonic waves to bats, dogs, dolphins and whales?
4. What is magnetostriction effect?
5. Which materials exhibit magnetostriction effect?
6. What is the formula for frequency of LC oscillator?
7. What is resonance? How does it take place? What is the advantage of using resonance in ultrasonic generators?
8. What do you mean by natural frequency? What is the formula for the natural frequency of vibrations of quartz crystal?
9. Explain with the help of necessary circuit diagram, how magnetostriction effect is used to produce ultrasonic waves. What are the advantages and disadvantages of magnetostriction oscillator?
10. What is piezoelectric effect? What is the difference between the direct and converse piezoelectric effect?
11. Which materials exhibit piezoelectric effect?
12. Draw a neat diagram showing the symmetry of the quartz crystal. How mechanical, electrical and optic axes are identified?
13. With the help of necessary diagram, explain how piezoelectric oscillator is used to produce ultrasonic waves. What are the advantages and disadvantages of piezoelectric oscillator?
14. What is the role of LC circuit in piezoelectric/magnetostriction oscillator?
15. What is the role of triode valve in piezoelectric/magnetostriction oscillator?
16. The range of frequencies of the ultrasonic waves that can be generated by piezoelectric oscillator is higher than that can be produced by magnetostriction oscillator. Why?
17. Magnetostriction effect is exhibited by only ferromagnetic materials. Why?
18. Discuss various methods used for detecting ultrasonic waves.
19. Explain how piezoelectric method can be used to detect ultrasonic waves.
20. Explain in detail, how a SONAR works. What are the applications of SONARs?
21. What is the difference between ECHO SONAR and DOPPLER SONAR?
22. Why RADAR cannot be used instead of SONAR? Why SONAR cannot be used instead of RADAR?
23. What is the equation used for calculating the velocity of ultrasonic waves through sea water?
24. Explain how ultrasonic waves are used for measuring thickness. What is the advantage of ultrasonic method for measuring thickness?
25. Explain how ultrasonic waves are used for nondestructive testing (NDT). What are the advantages and disadvantages of ultrasonic NDT?

26. Explain how ultrasonic waves are used in ultrasonography. What are the advantages and limitations of ultrasonography?
27. In which of the media such as gases, liquids and solids, the velocity of ultrasonic waves is maximum? Why?
28. Explain how ultrasonic waves are used for nondestructive testing. What are the advantages and disadvantages of ultrasonic nondestructive testing?
29. Explain how ultrasonic blood flow meters work.
30. Enlist any five distinct medical applications of ultrasonic waves.
31. What is Doppler effect. Does any phenomenon in daily life demonstrate the Doppler Effect?
32. Doppler Effect, in another sense also plays a role in astronomy. What is it?

REFERENCE BOOKS

1. Fundamentals of Physics, 9th Edition, Extended, Wiley Resnick, Halliday, Walker,
2. Introduction To Acoustics by Robert D Finch, Pearson India, 2016
3. Fundamentals of Acoustics, 4th Edition, Lawrence E. Kinsler, Austin R. Frey, Alan B. Coppens, James V. Sanders, 1999, Wiley
4. Fundamentals and Applications of Ultrasonic Waves, Second Edition, J. David N. Cheeke, CRC Press, Taylor and Francis Group
5. Understanding Ultrasound Physics, 3rd Edition, Sidney K. Edelman, E. S. P. Publishers, 2004

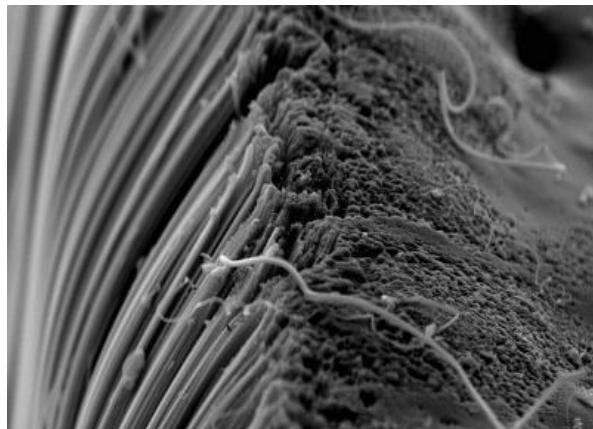
WORLD WIDE WEB

1. Ultrasonics Society of India, www.ultrasonicsindia.org
2. Ultrasonics (Journal), Elsevier, <https://www.journals.elsevier.com/ultrasonics>
3. <http://hyperphysics.phy-astr.gsu.edu/hbase/Sound/usound.html>

CHAPTER 8

Basics of Quantum Mechanics

(Wave-Particle Duality)



The photograph on left an electron microscope; a microscope with incredible resolving power and the photograph on right shows carbon nanotubes being observed through electron microscope. The resolving power of electron microscope is a few lakh times greater than that of optical microscope. The carbon nanotubes, which are the part of nanotechnology, have many applications. Both electron microscope and carbon nanotubes are based on laws of Quantum mechanics. What is Quantum mechanics?

Answer to this question is in this chapter

Index

8.1 INTRODUCTION

Quantum mechanics is the science of atoms, molecules and photons and their interactions

8.2 ROLE OF QUANTUM MECHANICS IN THE TECHNOLOGY

Quantum mechanics plays a decisive role when technology becomes delicate

8.3 DUALITY OF RADIATIONS

Radiations have wave as well as particle-like character

8.4 WHY QUANTUM MECHANICS IS CALLED QUANTUM MECHANICS?

As it asserts about the quantization of the energy

8.5 DE BROGLIE'S HYPOTHESIS

Just like radiations, matter also has dual nature

8.6 EXPERIMENTAL PROOF OF DE BROGLIE'S HYPOTHESIS (Optional)

Electrons can be diffracted and their wavelengths can be measured

8.7 CHARACTERISTICS OF DE BROGLIE WAVES

De Broglie waves are probability waves

8.8 VELOCITY OF DE BROGLIE WAVES

De Broglie waves have two velocities, one is phase velocity and another is group velocity

8.9 HEISENBERG'S UNCERTAINTY PRINCIPLE

Wavelike properties of subatomic particles lead to an unavoidable uncertainty in determining their motion

8.10 EXPERIMENTAL PROOF OF UNCERTAINTY PRINCIPLE-I: SINGLE SLIT DIFFRACTION OF ELECTRONS

The width of slit affects the uncertainties in position and momentum in an opposite manner

8.11 EXPERIMENTAL PROOF OF UNCERTAINTY PRINCIPLE-II: GAMMA RAY MICROSCOPE

A subatomic particle cannot be observed without disturbing it

8.12 HEISENBERG'S UNCERTAINTY PRINCIPLE AND QUANTUM MECHANICS

Heisenberg's uncertainty principle is consistent with many predictions of quantum mechanics

8.13 PARTICLE IN A RIGID BOX: WITH FIRST PRINCIPLES

Restriction leads to quantization

8.1 INTRODUCTION

Quantum mechanics is the science of atoms, molecules, photons and their interactions

The nature around us is miraculous, complex and yet to be fully understood. However, according to a simplified and a basic approach, at mega as well as micro scale, it is the interaction between matter and energy which governs all the operations in the universe. What is the fundamental nature of basic building blocks of the matter and energy in the nature? How do they behave? How do they interact? The basic building blocks of nature are atoms, molecules and photons. Indeed the property of each and every unique element in the universe is governed by the structure of its atom and the aggregation of atoms. It is the periodic table of elements which governs the properties and applications of materials and machines. Indeed materials and machines cannot supersede the periodic table. For example, let us ask following questions to ourselves...

1. Why silicon and germanium are semiconductors?
2. Why lithium, beryllium, graphite, copper, gold, aluminum, silver; indeed all the metals behave like conductors?
3. Why diamond is an insulator? Why glass and plastics are insulators?
4. Why iron, cobalt and nickel are ferromagnetic? Why magnesium, molybdenum, and lithium, are paramagnetic? Why copper, silver, and gold, are diamagnetic? Why FeO is anti-ferromagnetic and why Fe_3O_4 (the natural magnet) is ferrimagnetic?
5. Why a few elements and compounds such as niobium, Nb_3Sn behave like superconductors at low temperatures?
6. Why certain ceramic oxides of copper (called cuprates) such as $\text{YBa}_2\text{Cu}_3\text{O}_7$ or La-Ba-Cu-O etc. behave like superconductors in the temperature range 91-125 K?
7. Why laser is an amplified and coherent light?
8. What is the underlying theory behind the functioning of high rank instruments and devices such as transistor, laser, scanning electron microscope (SEM), Scanning Tunneling Microscope (STM), tunneling devices such as tunnel diodes, Josephson junction etc. ?
9. What is the foundational theory behind rapidly upcoming technologies such as nanotechnology, photonics, spintronics, and molecular/organic electronics?

Well! these and several such questions related with the properties , and off course the applications, of several elements and their compounds cannot be answered, at least principally, without learning quantum mechanics. Indeed quantum mechanics resolves all such riddles at the most basic and fundamental level i.e. atom, its structure and it's aggregation with its neighboring atoms. The theory of quantum mechanics comprises of photons-the basic energy quanta and their interaction with the building blocks of matter i.e. the atoms.

Two aspects of quantum mechanics are important...one its evolution and another it's predictions. Quantum mechanics did not evolve in a planned manner and neither was its initial aim to understand the structure i.e. electronic configuration of atoms. Quantum mechanics began and proceeded as per the standard method of physics...i.e. asking the questions and trying to answer them...logically. In was the historical period of 1897 to 1940 where the ideas of quantum

mechanics evolved and settled. Many contemporary and Nobel prize winning Physicists namely Niels Bohr, De Broglie Werner Heisenberg, Erwin Schrödinger, Max Born, Wolfgang Pauli, Paul Dirac played an active role in the development of quantum mechanics. Interestingly, unlike in other cases, none of these Physicists was originally an engineer. It would have been impossible for engineers to participate in the development of quantum mechanics. This is simply because engineers believe too much in common sense. And quantum mechanics requires an adventure to jump beyond the boundaries of commons sense if it is permitted by rational thinking supported by the predictions of theories and the results of experiments. Indeed quantum mechanics requires an open mindedness to accept the nature as it is. Further, as far as predictions and their experimental verifications are concerned, quantum mechanics can be considered as the most winning theory, because almost all predictions of quantum mechanics have been experimentally verified.

8.2 ROLE OF QUANTUM MECHANICS IN TECHNOLOGY

Quantum mechanics plays a decisive role when technology becomes delicate

In spite of its incredible success in explaining matter and energy in terms of their basic building blocks such as atoms molecules and photons, many times, students of engineering and technological disciplines tend to ignore quantum mechanics by treating it as an abstract theory. This is due to a wrong perception. Quantum mechanics plays a decisive role when technology becomes delicate, i.e. devices become exceedingly small. Transistor (whose invention helped in reduction of size, cost, and power consumption of electronic gadgets and enhancement of their speeds), The high precision microscopes such as Scanning Electron Microscope, (SEM), Transmission Electron Microscope (TEM), Scanning Tunneling Microscope (STM), Atomic Force Microscope (AFM)which are playing a prime role in nanotechnology research) and laser (whose applications in CD ROMs and fiber optics boosted the IT revolution) are all based on quantum concepts. Moreover, an amazing device called Josephson junction which finds applications in SQUIDS (capable of sensing magnetic fields as small as 10^{-18} Tesla, such as those from heart and brain), SET (single electron transistor) is an outcome of typical effect called tunnel effect, which can be explained only by quantum physics. Further, the rapidly upcoming branches of Physics such as Photonics (photon-based-electronics), Spintronics (electronics based on spins of electrons that may give birth to Quantum Computer), Nanotechnology, Molecular electronics which promise to give a completely different and better shape to existing technologies, are also an outcome of our understanding of matter and energy, which would have not been possible without quantum mechanics.

8.3 DUALITY OF RADIATIONS

Radiations have wave as well as particle-like character

Though the attempts to understand the fundamental of nature of light (yet to be fully understood) and its related phenomena such as reflection, refraction, interference, diffraction, polarization, which formulated the optics date back to Newton and Young, Fraunhofer and Fresnel, Huygen and Rayleigh, it was James Clerk Maxwell who put forth the idea that light is a

form of electromagnetic wave...yes WAVE! Light is one of seven electromagnetic waves which exist in nature. The electromagnetic waves and their characteristics was predicted by four famous equations named Maxwell's equations.

The physicists before 1900 were quite happy about the wave theory of light and it's achievements. Indeed the "wave optics" which gave birth to several optical instruments from microscopes to telescopes and interferometers to diffractometer were the successful applications of the wave optics. However, the foundations of physics were again to be shaken. Max Planck (Nobel prize in Physics in 1918), became the first Physicist to introduce the idea about the quantization of the radiation, rather in an unconvincing manner. The experimentally observed black body spectrum was not being correctly and completely explained by using the wave theory of infrared radiation. The formula developed by considering the wave theory of radiation could explain only the limited part of the experimentally observed black body spectrum. The development of the formula which explained the spectrum completely and fully was based on a newly introduced idea of Max Planck that black body could emit and absorb the radiation, not continuously and in the form of waves, but in a quantized or particle-like manner. The quanta of energy absorbed emitted by black body had their energy given by

$$E_n = nh\nu \quad \dots(8.1)$$

and the basic quantum of the energy was given by

$$E = h\nu \quad \dots((8.2)$$

Max Planck, the originator of the quantum theory of radiation remained the first physicist to do so, but he was not the only one. Albert Einstein (Nobel prize in physics in 1921) proposed his famous equation of the Photoelectric effect using the concept of Planck's quanta. Another confirmation to the quantization of the energy occurred in Compton effect, where Arthur Holly Compton (Nobel prize in physics in 1927) observed that, X rays, when scattered through the carbon scatterer carried the wavelength larger than that of incident one. While wave theory completely failed to explain why X rays should carry higher wavelengths after scattering, the quantum theory provided a complete, adequate and accurate explanation. During scattering the X ray 'quanta' collide with the electrons in the scatterer and lose their energy if the collisions happen to be inelastic. This results in loss in frequency and thus, an increase in the wavelength.

The three major experiments such as interference, diffraction and polarization in optics convincingly declared that radiations have wave- like character. But the results of black body radiation, photoelectric effect and Compton effect, for their complete and unambiguous explanation, indicated that, at least in a few situations, radiation exhibit particle-like (quantum) properties. Let us first understand, what quantization means.

Fig (8.1) represents a simple sinusoidal progressive wave. Progressive waves are boundary-less waves. i.e. they are infinitely long in space. The simplest example of such a wave is the wave generated on infinitely long string. These waves are represented by following equation.

$$y = y_0 \sin(kx - wt) \quad \dots(8.3)$$

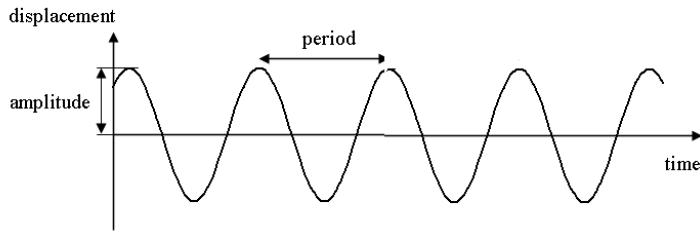


Figure (8.1)A single and infinitely long sinusoidal wave is not quantized

As the wave does not have any restriction in the space, it can be oscillated with any small, medium or large wavelength. An infinitely long string can be oscillated with any wavelength without any restriction. In another language, the wavelength (i.e. frequency, and hence the energy) of an infinitely long wave is not restricted. One of the meanings of quantization is ‘restriction’ and another meaning is ‘packetization’. Thus an infinitely long wave is not quantized.

Now let us consider waves bounded between boundaries (Fig 8.2). As we all know, such waves are bounded across rigid supports. Now here the wave cannot be generated in an unrestricted manner. Only those waves are possible which have zero displacement at the boundaries (these conditions are called as boundary conditions). This imposes restriction/quantization on the wavelengths of such waves. The restricted/quantized wavelengths of such waves are given by

$$\lambda_n = \frac{2L}{n} \quad \dots(8.4)$$

Where $n (\neq 0), 1, 2, 3, 4 \dots \dots$

Note that n is restricted to take only integer values. If n is allowed to take fractional values then , the corresponding waves will not satisfy the boundary conditions, where it is expected that a standing wave should have a zero displacement at the rigid boundaries. Thus n is restricted/quantized, and thus can be called as **quantum number**. Interestingly n cannot take zero value, as otherwise the wavelength of the wave in the finite region would become infinite. Later on we will come to know that nonzero value of n is directly related to the concept of nonzero ground state energy, one of the basic concepts of Quantum mechanics. The restriction on the wavelengths of the waves leads to quantization of their energy. (wavelength is solely related to energy. For ex., a small wavelength corresponds to higher energy and a long wavelength corresponds to lower energy)

The figures (8.1) and (8.2) and the above discussion collectively lead to one of the most fundamental principles in quantum mechanics...

The energy of free wave is not quantized and energy of a bounded wave is always quantized.

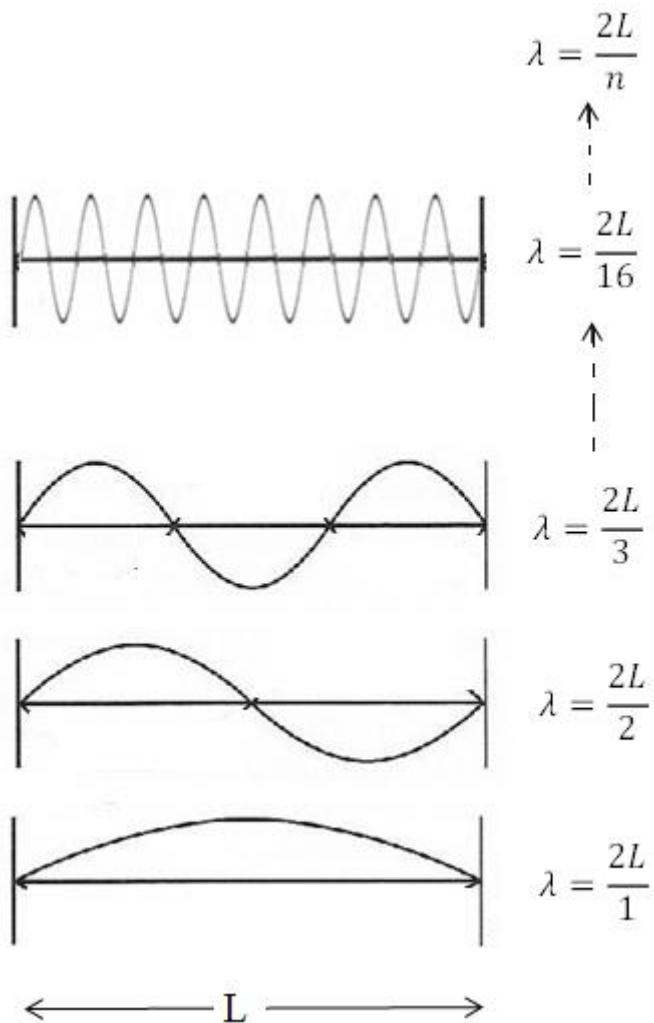


Figure 8.2 Standing waves (are quantized)

Thus restriction on the motion of anything...may it be wave or particle...leads to quantization. We will learn more details of this aspect in next chapter.

8.4 WHY QUANTUM MECHANICS IS CALLED AS QUANTUM MECHANICS?

As it asserts about the quantization of the energy

Consider an analogy. Suppose you want to purchase milk from a dairy. In some cases you may ask for any amount of milk, say 12.45 ml or 167.56 ml or so and purchase it. But as we know, the milk is generally not sold in this way. The milk is packetized. The packets are filled with either half a litre or one litre milk. Thus if we want to purchase milk, we can purchase it only in the integer multiples of half liter. This is quantization of milk. As an another analogy consider the distribution of the students in various divisions. In some institutes, the students are admitted to various divisions according to alphabetical order of their surnames. But in a few

colleges, the students are arranged in various divisions according to their merit. This is quantization of divisions. These analogies perfectly apply to the energy of radiation (and even matter also). The radiations consist of packets of discrete amounts of energy. Later on we will understand that this is true for atoms and molecules also. The energy levels of atoms and molecules (and radiations) are also discrete i.e. quantized.



Louis-Victor de Broglie (1892-1987), Although, a student of history, he made an important contribution in Physics also. Inspired by Niel's Bohr's theory of quantization of electron orbits, he thought that electrons must be having a sort of periodic properties. In his doctoral thesis in 1924, he proposed that moving bodies have wave-like properties that complement their particle properties. Later on his hypothesis was confirmed in electron diffraction experiments. His suggestion was followed by Erwin Schrodinger and others to develop quantum mechanics which explained wide variety of atomic phenomena. De Broglie was awarded a Nobel prize in 1929

8.5 DE BROGLIE'S HYPOTHESIS

Just like radiations, matter also has dual nature

Louis-Victor de Broglie, during his Ph.D. thesis made a bold suggestion that, if radiations have dual character, then analogically, matter may also have dual character. This meant that if, radiations which were proven to be waves at first, could exhibit particle like properties in some other experimental situations, then the material particles which are known to exhibit particle properties, may also exhibit wave properties in some other experimental situations. As such, the particle properties of waves and wave properties of particles are just beyond our day-to-day i.e. classical common sense. We understand that a particle is localized, it has mass, momentum, kinetic energy and definite position, while a wave is delocalized i.e. spread in space. Further, it has amplitude, wavelength, frequency, period and displacement etc. Moreover, a wave does not have a definite position in the space. How can the same quantity say, a radiation, or a material particle have both the diametrically opposite properties? Well! Such questions arise only if one thinks in a classical sense based on experience in the day to day life. In day to day life wave clearly appears like a wave and particle clearly appears as a particle. However, the entities which are beyond the perceptions of human senses cannot be assigned a definite nature, just by 'looking' at them. Let us accept that the nature (i.e. whether a wave or particle) of none of the radiations can be directly 'seen' by eye and neither be visualized using any microscope. Thus one has to believe on the results of experiments. The interference of light waves produces maxima and minima due to superposition of waves. This happens due to destructive or constructive interference of the waves. Now, the particles can collide with each other but can't interfere destructively or constructively. Thus the very fact that light generates interference pattern (or diffraction pattern) clearly indicates that light must be behaving like a wave at least in the interference, diffraction and polarization experiments. On the other hand, all the results of the photoelectric experiment conclusively indicate that, at least in the photoelectric effect experiment, light must be 'behaving' as if its energy is localized in the packet. (for example, in the photoelectric experiment the electrons cannot be instantly emitted if the energy of the light

does not fall on the metal ‘at once’ and in the form a packet/quantum/photon). What we need to perceive in Physics should depend upon the results of experiments and their logical interpretation and not on the common sense based on the sensory organs. It should be noted that radiations (and particles also) exhibit only one property i.e. wave or particle at a time.

The application of the dual nature to the matter by De Broglie was based on his (and every Physicist’s) belief that nature loves symmetry. What does this mean? Let us see a few examples...

1. Action and reaction are equal and opposite in direction
2. Time dependent magnetic field generates electric field and symmetrically time dependent electric field generates magnetic field
3. Our image in the mirror is (anti) symmetric
4. An electron has antiparticle, symmetrically a proton also has an antiparticle
5. Principle of least action in mechanics and principle of least time in optics are equivalent

Thus according to De Broglie, matter and to be more precise, every material particle has wave-like properties. In another language, a certain kind of wave is associated with every material particle, may it be an electron or a speck of a dust or even a cricket ball. These waves are called as ***De Broglie waves*** or ***matter waves***.

What is the wavelength of a De Broglie wave? This can be derived using the analogy with photon

For photon

$$E = pc \text{ and } E = h\nu = h\frac{c}{\lambda}$$

$$\text{Thus } h\frac{c}{\lambda} = pc$$

$$\Rightarrow \lambda = \frac{h}{p} \quad \dots(8.5)$$

The formula for wavelength in the above equation works for photon, but De Broglie, using an analogy, applied the same formula for the wavelength of material particle. In principle, De Broglie’s hypothesis works for all the material entities including electron, proton, atom, molecule, bullet, cricket ball and even a planet or its satellite. However, in our day-to-day life none of the material object appears to be behaving like a wave. Let us understand the reason behind this through an example

Example (8.1):

Calculate the De Broglie wavelength of the (a) electron moving at 2×10^6 m/s and a cricket ball of mass 200gm moving at 20 m/s. Which of this entity particle behaves more like a wave and which of the entity behaves more like a particle?

Solution:

For electron

$$\begin{aligned}\lambda &= \frac{h}{p} \\ \lambda &= \frac{h}{mv} \\ \lambda &= \frac{6.63 \times 10^{-34}}{9.1 \times 10^{-32} \times 2 \times 10^6} \\ \lambda &= 3.64 \times 10^{-10} m \\ \lambda &= 3.64 \text{ A}^\circ\end{aligned}$$

For cricket ball

$$\begin{aligned}\lambda &= \frac{h}{p} \\ \lambda &= \frac{h}{mv} \\ \lambda &= \frac{6.63 \times 10^{-34}}{200 \times 10^{-3} \times 20} \\ \lambda &= 1.6575 \times 10^{-10} m \\ \lambda &= 1.6575 \times 10^{-34} m\end{aligned}$$

Let us note two observations. The De Broglie wavelength of electron, though small w.r.t. our day to day standards, is fairly comparable with its own size (10^{-16} m) and quite comparable with the size of the region (i.e. atom or a molecule) in which it exists. Further, like X rays, the wavelength of this order can be easily measured using electron diffractometers. Thus electron certainly behaves like a wave.

This is not true for cricket ball, its wavelength is extremely small as compared to its size as well as the size of the region in which it exits (i.e. cricket ground). Further, an experimental set up to measure such a small wavelength is yet to be invented. Thus cricket ball, though principally a wave, appears like an object.

Thus results of this problem lead to an extremely important principle of quantum mechanics, which asserts that the wave-like properties are more conspicuous in case of only subatomic entities.

Example (8.2):

Calculate the De Broglie wavelength of the cricket ball in Example (8.1) if we were living in

universe where Planck's constant was 6.63 J-sec. Discuss the consequences of the results of this problem.

$$\begin{aligned}\lambda &= \frac{h}{p} \\ \lambda &= \frac{h}{mv} \\ \lambda &= \frac{6.63}{200 \times 10^{-3} \times 20} \\ \lambda &= 1.6575m\end{aligned}$$

It can be observed that now the De Broglie wavelength of cricket ball is considerable and also comparable with its size. It is measurable also. Thus, in such universe, the cricket ball will not be observed like an object but like a wave. But what kind of wave it will appear like? Wait! This is still to be learnt.

De Broglie wavelength in terms of other physical parameters

We know that

$$\begin{aligned}K.E. &= \frac{1}{2}mv^2 = \frac{p^2}{2m} \\ P &= \sqrt{2mK}\end{aligned}$$

Thus

$$\lambda = \frac{h}{\sqrt{2mK}} \quad \dots(8.6)$$

If a charged particle of charge q and mass m is accelerated with a potential difference V , then the electrostatic work done on the charged particle is converted into its kinetic energy and is given by

$$\frac{1}{2}mv^2 = qV$$

Thus from Eqn. (8.6)

$$\lambda = \frac{h}{\sqrt{2mqV}} \quad \dots(8.7)$$

Eqn. (8.7) holds for any charged particle of charge q and mass m . Thus the De Broglie wavelengths of charged particles like proton, deuteron, and alpha particle can be calculated. Eqn. (8.7) can be made more specific for electron by substituting the values of charge and mass of electron and the Planck's constant. On doing the calculations and expressing the De Broglie wavelength in A° .

$$\lambda = \frac{12.26}{\sqrt{V}} A^\circ \quad \dots(8.8)$$

Let understand a few other important concepts of Quantum mechanics by solving a few more problems.

Example (8.3):

Calculate the wavelengths of photons of energies 1 eV, 1 keV and 1 MeV. Comment of the results

For photons

$$\begin{aligned} E &= h\nu = h\frac{c}{\lambda} \\ \lambda &= h\frac{c}{E} \\ \lambda &= 6.63 \times 10^{-34} \frac{3 \times 10^8}{1.6 \times 10^{-19} \times E} \times 10^{10} A^\circ \\ \lambda &= \frac{12431}{E(eV)} A^\circ \end{aligned}$$

For 1 eV photon $\lambda = 12431 A^\circ$

For 1 KeV photon: $\lambda = 12.431 A^\circ$

For 1 MeV photon $\lambda = 0.12431 A^\circ$

Thus amongst the electromagnetic radiations, low energy radiations like radio waves (1 eV) have larger wavelengths and the high energy radiations like gamma rays (1 MeV) have considerably smaller wavelength. Indeed one may be tempted to conclude that as one move from radio waves to gamma rays, the ‘waveness’ of the radiation decreases and its ‘particleness’ increases.

Example (8.4):

Calculate the De Broglie wavelengths of 1 keV photon and 1 keV electron. Compare them and interpret the results

For 1 keV photon, as per the previous problem $\lambda = 12.431 A^\circ$

Now for 1 keV electron

$$\lambda = \frac{h}{\sqrt{2mK}}$$

$$\lambda = \frac{6.63 \times 10^{-34}}{\sqrt{2 \times 9.1 \times 10^{-31} \times 1000 \times 1.6 \times 10^{-19}}}$$

$$\lambda = 3.88 \times 10^{-11} m$$

$$\lambda = 0.388 \text{ A}^{\circ}$$

Thus the De Broglie wavelength of a 1keV electron is much smaller than the photon of the same energy. Thus one may be tempted to conclude that, amongst the photon and electron of same energy, photon has more '*waveness*' and less '*particleness*' and electron has less '*waveness*' and more '*particleness*'

Example (8.5):

Calculate the energy of electron and photon both having wavelength 1 A[°]

Solution:

We have for electrons

$$\lambda = \frac{h}{\sqrt{2mK}}$$

$$\lambda^2 = \frac{h^2}{2mK}$$

$$K = \frac{h^2}{2m\lambda^2}$$

$$K = \frac{(6.63 \times 10^{-34})^2}{2 \times 9.1 \times 10^{-31} \times (10^{-10})^2}$$

$$K = \frac{4.40 \times 10^{-67}}{1.82 \times 10^{-50}}$$

$$K = 2.42 \times 10^{-17} J$$

$$K = \frac{2.42 \times 10^{-17} J}{1.6 \times 10^{-19} J/eV}$$

$$K = 151 eV$$

and for photons

$$E = h\nu = h \frac{c}{\lambda}$$

For Photon

$$E = 6.63 \times 10^{-34} \frac{3 \times 10^8}{10^{-10}}$$

$$E = 1.989 \times 10^{-15} J$$

$$E = 12431 eV$$

Thus for possessing same wavelength the photon has to carry much more energy than an electron.

Example (8.6):

Use De Broglie's hypothesis to prove that electron cannot exist inside the nucleus

Solution

If electron had existed inside the nucleus then its maximum De Broglie wavelength would not exceed the size of nucleus i.e. $10^{-14} m$

Thus for $\lambda_{max} = 10^{-14} m$

$$P_{min} = \frac{h}{\lambda_{max}}$$

$$P_{min} = \frac{6.63 \times 10^{-34}}{10^{-14}}$$

$$P_{min} = 6.63 \times 10^{-20} kgm/sec$$

$$mv_{min} = 6.63 \times 10^{-20}$$

$$v_{min} = \frac{6.63 \times 10^{-20}}{m}$$

$$v_{min} = \frac{6.63 \times 10^{-20}}{9.1 \times 10^{-31}}$$

$$v_{min} = 7.28 \times 10^{10} > c, \text{Speed of light}$$

Thus if electron had existed inside the nucleus, its minimum velocity would exceed the speed of light. This would violate the special theory of relativity. Thus electron can't exist inside the nucleus

OR

$$P_{\min} = 6.63 \times 10^{-20} \text{ kg m/sec}$$

$$E_{\min} = \frac{P_{\min}^2}{2m}$$

$$E_{\min} = \frac{(6.63 \times 10^{-20})^2}{m}$$

$$E_{\min} = \frac{4.39 \times 10^{-39}}{9.1 \times 10^{-31}}$$

$$E_{\min} = 4.8 \times 10^{-9} \text{ J}$$

$$E_{\min} = \frac{4.8 \times 10^{-9} \text{ J}}{1.6 \times 10^{-19} \text{ J/eV}}$$

$$E_{\min} = 3.02 \times 10^{10} \text{ eV}$$

$$E_{\min} = 30190 \text{ MeV} \gg 8.8 \text{ MeV} (\text{the maximum B.E. of the nucleus})$$

Thus if electron had existed inside the nucleus, its energy would be far greater than maximum binding energy of the nucleus. Thus nucleus can never trap an electron.

De Broglie's hypothesis has one more interesting consequence. It does not allow a subatomic particle to take rest. This is because if particle takes a rest its momentum p will be equal to zero and wavelength λ will be equal to ∞ . For any subatomic particle existing in finite regions such as atoms, molecules and solids, the De Broglie's wavelength cannot be infinite, in fact it cannot be greater than the size of the region in which it exists. This indicates that, due to the constraint that its De Broglie's wavelength to be; all subatomic particles have to be restless. Thus the subatomic world (or so to say quantum world) is restless. In the language of quantum mechanics, the **ground state energy** (minimum possible energy) of any subatomic particles cannot be reduced to zero.

Example (8.7):

De Broglie's hypothesis suggests that material objects have to be restless. How, then the objects in our day to day life can be at the rest?

Solution

Let us solve this problem by assuming a suitable data. Consider a cricket ball of mass 0.5 kg existing in a room of length 10 m. The maximum De Broglie wavelength of such cricket ball can be $\lambda = 10$ m. The corresponding minimum momentum is then

$$p = \frac{h}{\lambda}$$

$$p = \frac{6.63 \times 10^{-34}}{10}$$

$$p = 6.63 \times 10^{-35} \text{ kg m/s}$$

This momentum is too small to be considered. As can be noticed, this is due to an extremely small value of the Planck's constant. What would happen, if Planck's constant possessed a different value?

Example (8.8):

What would be the minimum momentum of the cricket ball in the above problem, if Planck's constant were 6.63 J.s?

Solution:

The calculation shows that the minimum momentum would be $66.3 \text{ kg } \frac{\text{m}}{\text{s}}$. This suggests that all the objects in our daily life would be restless if Planck's constant were really 6.63 J.s. But nature has cleverly chosen an appropriate and an extremely small value of Planck's constant to avoid the complications in our life! Extremely small value of Planck's constant allows us to take rest!

Example (8.9):

Calculate the energy of electron when it is accelerated by a potential difference of 100 KV

Solution:

We have, for an electron,

$$\lambda = \frac{12.26}{\sqrt{V}} \text{ A}^\circ$$

$$\lambda = \frac{12.26}{\sqrt{1000000}} \text{ A}^\circ$$

$$\lambda = 0.039 \text{ A}^\circ$$

Thus, if electrons are accelerated at 100 KV, their De Broglie wavelength becomes 0.039 A° . This wavelength is approximately 1.25 lakh times smaller than the wavelength of light. According to the theory of diffraction, a general formula for the resolving power of a microscope (or a telescope) is $R.P. = \frac{d}{1.22\lambda}$. Thus, smaller the wavelength, higher is the resolving power. Thus the resolving power of electron microscope is almost 1.25 lakh times greater than that of optical microscope. This is what makes electron microscope incredibly better than the

optical microscope. Electron microscopes are routinely used to observe the tiny entities like microorganisms, integrated circuits. Electron microscopes are invariably used in nanotechnology. The nano-devices or nanomaterials whose size is typically in a range of a few nanometers are beyond the reach of optical microscope. But these can be easily observed by electron microscope. In next chapter we will learn Scanning Tunneling Microscope (STM), which is still better than electron microscope.

The wavelike properties of electrons also make one more powerful device possible. This is electron diffractometer. We know Bragg's X ray spectrometers. These are routinely used in X ray crystallography and X ray spectroscopy. The crystalline properties of materials can be identified by analyzing their X ray diffraction patterns. In the same manner, the electron diffraction pattern of a material provides additional insights into the crystal structure of the materials.



Ernst August Friedrich Ruska (1906–1988): He was a German Physicist, who shared a Nobel Prize in Physics in 1986 for constructing world's first electron microscope. He received his education in Technical University in Germany. In 1933, he proposed that as the resolving power of an optical microscope is limited by wavelength of the light (approximately 5500 \AA), an electron microscope, based on wave properties of electron could have much greater resolving power, as electron's wavelength quite less than that of light. He also constructed a magnetic lens, which could magnify an electron image; much in a similar manner an optical lens magnifies the optical image. He then constructed electron microscope in 1933, the same year in which he received his Ph.D. Electron microscope was immediately put in use for medical and biological applications by his brother Helmut. Ruska's initial career was in Siemens but, later on, Ruska worked till his retirement, as a Professor in the same Technical University, where he was educated.



World's first electron microscope constructed by Ernst Ruska in 1933

8.6 DE BROGLIE'S HYPOTHESIS AND BOHR'S MODEL

De Broglie's Hypothesis is consistent with quantization of Bohr orbits

De Broglie, who proposed his hypothesis in 1924 could logically explain that, if electrons were treated as waves instead of particles, then the quantization of the orbits in Bohr's model could be logically explained. Let us see how.

As we know, in 1913 Niels Bohr proposed a typical model of the hydrogen atom, with following postulates

1. The orbits of electrons in the atom are quantized. Only those orbits are possible, where the angular momentum, radius and the energy is given by

$$L = n \frac{h}{2\pi} = n\hbar$$

$$r = 0.511n^2 A^\circ$$

And

$$E_n = -\frac{13.6}{n^2} \text{ eV}$$

2. The second postulate asserts about emission of a quantum of energy (photon), when electron makes transition from upper to lower orbit.

Thus the orbits of electrons in hydrogen atom (and as you will further learn...in any atom) are quantized. Though this model correctly explains the observed spectrum of the hydrogen atom, the model is empirical, i.e. there is no consistent explanation to the fundamental questions such as why electron rotates only certain allowed orbits and why not in other orbits. Further the model also does not explain why electron does not emit radiation in stationary-allowed orbit.

However, the Bohr's model receives a logical support in the light of De Broglie's hypothesis, which treats electrons as waves

If electron, as a wave is to be smoothly fitted in the orbit, then following relation must be satisfied

$$\text{Total length of the orbit} = \text{circumference} = 2\pi r = n\lambda \quad \dots(8.9)$$

Refer Fig. (8.3): Only complete waves can be smoothly fitted in the orbits, as after curling a complete wave (length = $n\lambda$) in the orbit, the point of zero displacement will smoothly match with the point of zero displacement. The point where the start and end of the wave meets will then have only one displacement at a point. In case of incomplete waves (length $\neq n\lambda$) such as those shown in the second part of Fig. 8.3, when such wave is curled, the point where these waves joint will have two displacements and thus the wave fitting will not be smooth. In another language, only those orbits are possible for which Eqn. (8.9) is satisfied. This logically explains why Bohr orbits are quantized. Indeed, it was 1913 when Bohr applied the idea of quantization of radiation to the matter i.e. electron orbits in Hydrogen atom, and then it was in 1923 that, by

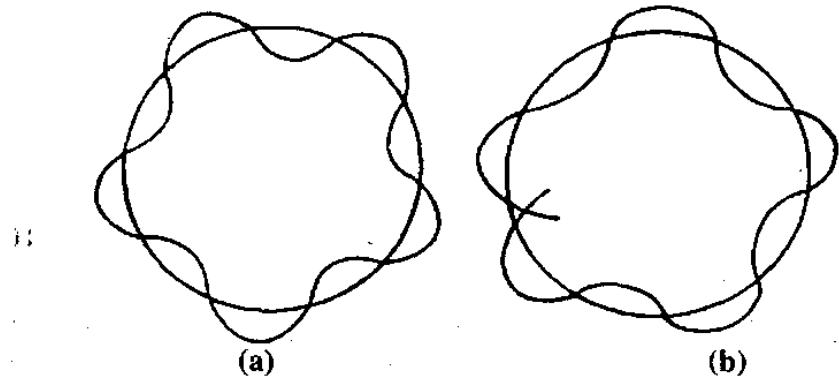


Figure (8.3) Waves satisfying Eqn. (8.9) can only be smoothly fitted in Bohr orbit. Only such orbits are possible, others are forbidden

assigning the wave properties to the electrons, the quantization of Bohr orbits was fitted in a logical and sound theoretical framework. So in physics also there are chances of coincidences!

Let us go one step ahead from Eqn. (8.9). Substituting $\lambda = \frac{h}{p}$ in Eqn. (8.9) we get

$$2\pi r = n \frac{h}{p}$$

Rearranging

$$rp = n \frac{h}{2\pi}$$

Thus angular momentum $L = n\hbar$

This gives the condition for existence of Bohr's allowed orbits which Bohr had empirically suggested. This confirms that the idea of quantization which was first applied to radiation, smoothly and logically extends to matter also. Further, this discussion also gives a confirmation to De Broglie's hypothesis.



Clinton Joseph Davisson (1881–1958): He was an American physicist, who received a Nobel prize in Physics in 1937 for the discovery of electron diffraction in his famous Davisson-Germer experiment. This experiment proved it without any doubt that electrons have wave-like properties. This Nobel prize was jointly awarded to Davisson and G. P. Thomson who had also worked on electron diffraction during the same period. Davisson received his degree education in University of Chicago and Ph.D. in Princeton University. After receiving Ph.D. he made his career in Carnegie Institute of Technology, Bell Telephone Laboratories and finally at the University of Virginia. In 1927, while working in Bell labs, along with his colleague Lester Germer, he showed that electrons could produce well defined diffraction pattern if they were passed through nickel. In this experiment they also calculated the wavelength of an electron which correctly matched with that calculated by using De Broglie formula. The crater Davisson on the Moon is named after him.

Sir George Paget Thomson (1892-1975): He was an English Physicist, who shared a Nobel prize in Physics, along with Davisson for proving wave properties of electron in electron diffraction experiment. Notably, he was the son of a Nobel laureate, sir J. J. Thomson, who received it for discovery of the electron (as a particle!). He received his education in the famous Trinity College (London). Imperial college was one of the institutes, where he made his career also. Like many contemporary physicists, he was also associated in making nuclear bomb. In fact in the later part of his life, he worked as a Nuclear Physicist. It is a unique coincidence that, J. J. Thomson and G. P. Thomson, both of whom won Nobel prizes in 1906 and 1937, revealed the particle and wave aspects of the same entity named electron. Electron diffraction does not just prove the wave properties of electron, but it also forms the basis of a powerful analytical technique named electron diffraction

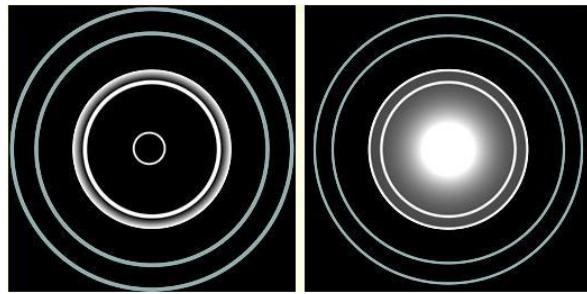


8.7 AN EXPERIMENTAL PROOF TO THE DE BROGLIE'S HYPOTHESIS(OPTIONAL)

Electrons can be diffracted and their wavelength can be measured

In 1927, Davisson and Germer and G.P. Thomson independently carried electron diffraction experiments in USA and UK. Davisson and Germer diffracted the electrons thorough Nickel crystal and found that electrons, after scattering through Bragg planes of the Nickel crystal, produced well defined diffraction patterns. By measuring angle of diffraction (θ) and substituting it in the Bragg's relation ($2dsin\theta = n\lambda$), they found that this wavelength was 1.67 \AA and it correctly matched with the wavelength calculated from theoretical relation $\lambda = \frac{12.26}{\sqrt{V}}$. In G.P.

Thomson's experiments also electrons exhibited well defined diffraction patterns after passing through Gold foil. Davisson, Germer and G.P. Thompson received Nobel Prize in physics in 1937.



Well defined electron diffraction patterns, as observed by G. P. Thomson

8.8 CHARACTERISTICS OF DE BROGLIE WAVES

De Broglie waves are the probability waves

We certainly know three kinds of waves...(i) sound waves (ii) light waves and (iii) waves on the water surface. We know almost everything about these waves, the quantities oscillating in them, their speeds and their other properties. For example, the quantity whose oscillations make up sound waves is pressure and sound waves travel with 340 m/s in air. The quantities whose oscillations make up light waves are electric and magnetic field. The light travels in vacuum with 3×10^8 m/s. In case of the waves generated on the water surface it is the oscillations of the molecules on the water surface which make up the wave. As we experience, sound waves can be ‘listened’ and light waves can be ‘seen’. Thus both these waves are physically detectable. Can we raise and then answer such questions for De Broglie waves? What is the quantity whose oscillations make up the matter waves? What is the speed with which De Broglie waves travel? Are De Broglie waves physically detectable? How do we realize their existence? Let us see how these questions can be answered.

Let us choose a notation for representing the oscillations of the De Broglie waves. Let this notation be ψ (pronunciation: psi). Now the most basic question is, what the physical significance is of ψ ? And more importantly, can ψ be experimentally measured?

Just as De Broglie did, we will choose the analogy of radiation, say light. Though light contains oscillations of both electric and magnetic field, it is electric field ‘ E ’ which is considered more significant (because light interacts with matter due to its electric field). Now we agree that light has dual nature as well as particle/photon nature. Let us now consider a situation, which we will try to describe by using both particleas well as wave properties. We know that if light is passed through a slit having its width comparable with its wavelength, then it produces a diffraction pattern containing maxima and minima. This diffraction pattern is shown in the Fig. (8.4).In the wave picture of light, the points of higher intensity correspond to higher values of E^2 (as intensity is proportional to the square of the amplitude). The points of lower intensity correspond to points of lower values of E^2 . Now, how we will describe the same situation in

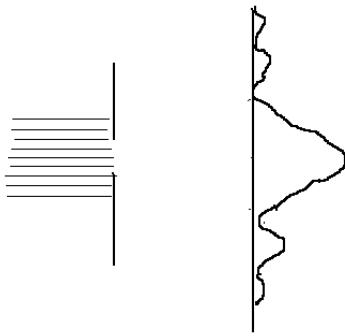


Figure (8.4): Diffraction pattern of light (or electrons)

terms of photon picture of light? Even if we treat light as a stream of photons, when we pass the light through a slit, ‘something’ will happen but the diffraction pattern will still be observed. In the context of photons the points of larger intensity will correspond to greater concentration of photons i.e. greater photon density i.e. greater number of photons per unit volume. The points of lower intensity will correspond to lesser concentration of photons i.e. lesser photon density i.e. lesser number of photons per unit volume. Thus, analogically, larger value of E^2 in the wave picture of light is equivalent to larger number of photons per unit volume in the photon picture and smaller values of E^2 in the wave picture is equivalent to lesser number of photons per unit volume. Now what will happen if the intensity of the light passing through the slit is decreased? This will result in the weakening of the diffraction pattern, but still the intensity at any point will be equivalent to E^2 in the wave picture and number of photons per unit volume in the particle picture. Let us extend this situation to its extreme and let us make the light passing through the slit so weak that it contains only one photon. The diffraction pattern, howsoever weak it may be, will still be produced. The points of maximum intensity will still correspond to maximum value of E^2 , but what about the photon picture? Can the intensity be now related with the number of photons per unit volume? Certainly not, as the entire light consist of only one photon and the entire diffraction pattern is made up of only one photon. Now the photon cannot be distributed or ‘fractionated’ over the entire diffraction pattern. The points of larger E^2 can neither be made equivalent to a given ‘amount’ of photon per unit volume, as photon is one individual entity. The most appropriate description in this case will be to relate E^2 to the probability of finding photon per unit volume. The points of larger intensity i.e. points of larger E^2 in this case are equivalent to the points where there is greater probability of finding the photons per unit volume and the points of lesser intensity will correspond to lesser E^2 i.e. lesser probability of finding the photon per unit volume. Now if the intensity of light is increased then obviously photons will acquire those places where their probability of occurrence is more. Then E^2 will still be equivalent to number of photons per unit volume. Thus fundamentally

Thus, for light, $E^2 \Rightarrow$ probability density of photons \Rightarrow probability of finding the photons per unit volume

Now nature loves symmetry. So let us pick up the same series of arguments and apply those to electrons. After passing electrons through a slit comparable to their wavelength, the electrons

will also produce diffraction pattern. And applying all the above set of arguments to electrons our final conclusion will be

For electrons, $|\psi|^2 \Rightarrow$ probability density of electrons \Rightarrow probability of finding the electrons per unit volume. This probabilistic interpretation of $|\psi|^2$ was given by Max Born (Nobel prize shared in Physics in 1954). Thus the De Broglie waves are neither the pressure waves and nor the electromagnetic waves; they are probability waves

Let us make one point clear here. A De Broglie wave described by ψ itself does not represent a probability wave. There are two reasons behind this. One is that when ψ is plotted, the corresponding De Broglie wave will have oscillations on positive as well as negative side. But probability cannot be negative. Further, in many cases the wavefunction ψ maybe a complex function (containing imaginary number i) and therefore cannot have real significance. Thus it is not ψ , but it is $|\psi|^2$, which represents the probability density of electron and more correctly any subatomic particle.

Let us make one more point clear. Suppose that the value of $|\psi|^2$ at a given point is 0.5. Thus there are 50 % chances of finding the subatomic particle, say an electron, in a unit volume around that point. Thus, out of 10 experiments, 5 times the electron will not be found and 5 times it will be found there. Now when electron is not found it is completely absent and when it is found it is completely present. This means that 50% probability does not indicate chances of getting 50% of the electron, but it indicates 50% chance of getting the complete electron. Thus probability of event does not mean event itself. Or in another language, the electron itself is not distributed over the wave, but it is the probability of finding the electron which is distributed over the $|\psi|^2$.

Let us understand another interesting aspect of this discussion. Consider a De Broglie wave represented by ψ . In next chapter we will learn that such ψ for any subatomic particle can be obtained by solving Schrödinger's equation for the corresponding particle. Let us assume that the value of $|\psi|^2$ at a given point is 50% and at other points, say it is 20%, 10% etc. Naturally if we want to detect an electron, we will choose the point of maximum probability i.e. 50%. Thus out of 10 experiments, 5 times we have a chance of getting the electron there and in 5 cases there is no chance. Suppose we start performing the experiments. Assume that we fail in detecting the electron in the first two experiments, but succeed in the third one. In the third experiment we will detect electron as a complete particle having charge equal to 1.6×10^{-19} C and mass equal to 9.1×10^{-31} kg. Further, as electron is detected, there is no need of paying attention to rest of the probability wave. Thus we need the De Broglie wave before we detect electron, and after detecting the electron we do not need it. Thus as one famous Nobel Prize winning Physicist, William Bragg puts it, '***everything in future is a wave and everything in past is a particle!***'! But can electron ever be detected at a fixed position? No!, Never! This is because, for detecting the electron, we will have to illuminate it with light, and light, even if it consists of a single photon, will kick the electron away from its original position in an unpredictable way. The electron can never be described in the language of 'certainty'; its description needs language of probability. Thus electron is both wave and particle i.e. it is a '***wavicle***'. We will elaborate these concepts in a greater depth in the next chapter.

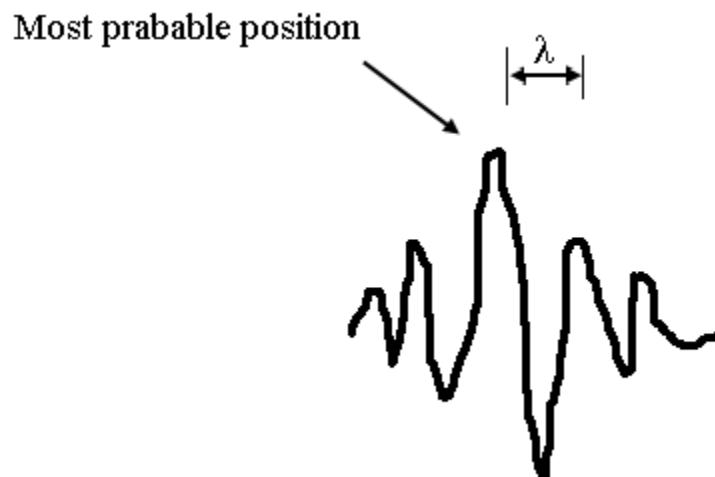
Let us now learn a few more characteristics of the De Broglie waves. There are two kinds of particles...one is an absolutely free particle and another is the particle confined in a definite space. Absolutely free particle is entirely an idealization. Almost all the subatomic particles or

the objects that we encounter in day to day life are confined to a fixed region. So to say, electron is confined to atom, atom is confined to a molecule, a proton is confined to a nucleus and a cricket ball is confined to a box in which it is kept.

The free particle can exist anywhere in space and probability of finding such particle is same in the entire infinite space. The De Broglie wave associated with such particle should be a single progressive wave accompanying the entire infinite space. The amplitude of this wave should be same everywhere, as the probability of finding the corresponding particle is same everywhere. Further, this infinitely long wave does not have to satisfy any boundary condition and therefore it is not quantized (as we know, the waves generated on infinitely long string can have any wavelength...thus no restriction and hence no quantization). But, what about confined particles? De Broglie waves associated with such entities cannot be infinitely long progressive and un-quantized waves. But such waves should be limited to the size of the region in which their particle exists. For ex. the De Broglie wave associated with the electron in the atom can be maximally as long as atom itself. Such waves of limited length are not progressive waves but they are wave-groups. A wave-group is obtained by superimposing different progressive waves of different wavelengths. The infinitely long progressive wave and a limited wave-group are represented by following figures.



(a)



(b)

Figure (8.5) (a) A single progressive infinitely long progressive wave and (b) a wave-group limited in space

As it can be seen, the progressive wave has same amplitude everywhere, indicating that the probability of finding an unconfined and absolutely free particle is same everywhere. However this is not true for wave-groups. Wave-groups have different amplitudes at different points indicating different probabilities of finding a confined particle at different points. Further, as De Broglie wave-group is limited to the region of existence of particle, the probability of finding the particle beyond the region of confinement is zero.

In next chapter we will learn that the wave-groups are confined and therefore they do not represent a free progressive wave; they a standing wave confined to the boundaries of the region. Further we will also learn that such wave-groups have to satisfy various boundary conditions which are satisfied for only certain energies and therefore their energies are quantized. The quantization of energy of confined waves can also be understood from Fig 8.2, where standing waves on a string attached with rigid supports are observed. Such waves can be oscillated only with specific wavelengths. Thus their wavelength is quantized. Now, if these waves are considered as De Broglie waves then according to De Broglie's hypothesis (eqn . 8.6), their momentum is quantized. Thus the energy of a confined wave is quantized. The wave associated with absolutely free particle is infinitely long and has no restriction on wavelength, momentum and energy. Thus energy of a free particle is not quantized. (*It is a fundamental principle in quantum mechanics that energy of a free particle is not quantized but energy of a bounded particle is quantized.*)

The conclusions of above discussion are as follows

1. De Broglie waves are probability waves
2. De Broglie waves associated with confined particles are wave-groups
3. The subatomic particles are restless
4. The energies of the De Broglie waves associated with the confined particles are quantized

Let us now understand one more aspect of De Broglie waves. Light travels in vacuum with 3×10^8 m/s. Sound travels in air with 340 m/s in air. What is the speed of De Broglie waves? As De Broglie waves are associated with moving material particles, one can logically expect that, as De Broglie wave has to accompany the particle with which it is associated, it should move with the same speed as that of particle. This seems to be obvious, but we will have to undergo a complex mathematical derivation to arrive at this conclusion. Let us see how.

8.9 THE VELOCITY OF DE BROGLIE WAVES

De Broglie waves have two velocities, one is the phase velocity and another is the group velocity

We know that in almost all cases the particles with which De Broglie waves are associated are confined and therefore De Broglie waves associated with such confined particles are wave-groups. For the while, let us consider an extreme case. Let us consider a free progressive and sinusoidal wave associated with an absolutely free particle. A simple progressive wave is represented by

$$\psi = \psi_o \sin(\omega t - kx) \quad \dots(8.9)$$

Where ω = angular velocity = $2\pi\nu$ and k = wavenumber = $\frac{2\pi}{\lambda}$

The phase velocity of such wave is defined as

$$u = \frac{\omega}{k}$$

$$u = \frac{2\pi\nu}{2\pi/\lambda}$$

$$u = \lambda\nu \quad \dots(8.11)$$

We have

$$\lambda = \frac{h}{p} \quad \dots(8.12)$$

Further, electron being a wave-particle, both equations of energy i.e. $E = mc^2$ and $E = h\nu$ are valid for electron, thus

$$E = mc^2 = h\nu$$

Thus

$$\nu = \frac{mc^2}{h} \quad \dots(8.13)$$

Substituting λ and ν in eqn. (8.11) we get,

$$u = \frac{h}{p} \times \frac{mc^2}{h}$$

$$u = \frac{h}{mv} \times \frac{mc^2}{h}$$

$$u = \frac{c^2}{v} = c \times \frac{c}{v} \quad \dots(8.14)$$

Now according to special theory of relativity v , the velocity of any particle/object cannot exceed c ($v < c$). Thus

$$u > c \quad \dots(8.15)$$

Does this mean that De Broglie waves travel faster than light? They certainly can't! , because then this will violate the theory of relativity. Let us recall that the above conclusion applies to an infinitely long sinusoidal progressive wave associated with an absolutely free particle. Such situation and therefore such De Broglie waves never exist. All objects/particles in the nature are confined and therefore the De Broglie waves associated with them are not infinitely long single and sinusoidal waves (which could move faster than light), but they are the wave-groups which are limited in space. Thus the fundamental question that we should ask is not... 'what is the velocity of De Broglie wave', but ... 'what is the velocity of De Broglie wave group'.

This derivation is optional. 

A wave-group is formed by superposition of several infinitely long sinusoidal (or cosine) progressive waves having marginally different wavelengths. The maximum number waves required to form a wave-group may tend to infinity (in such case the resultant wave-group becomes infinitesimally narrow), however, the minimum number of waves required to form a wave-group is two. Of course the wavelengths (and hence the wavenumbers and the angular velocity) of these two waves should be marginally different. This can lead to two cases. If the two superimposing waves of different wavelengths have same phase velocity, then the resulting wave-group will have same velocity as that of superimposing waves. However, if the phase velocities of the two superimposing waves (of different wavelengths) are different then the group velocity of the resulting wave-group will be different than the phase velocities of the superimposing waves. This is called as dispersion and we would like to focus on this case. In order to understand the nature of wave-group formed due to superimposition of two different waves of slightly different wavelengths (and hence frequencies), we may recall the phenomenon of beats. If sound waves of slightly different frequencies, say 330 and 332 Hz, but same amplitude (loudness) are superimposed then, as a result, we hear a sound wave of average frequency 331 Hz but of periodically varying amplitude, where the loudness varies as the difference of two frequencies. Let us consider the superposition of such two sinusoidal waves of angular frequencies ω and $\omega + \Delta\omega$ and wavenumbers k and $k + \Delta k$. These waves are traveling with different phase velocities.

$$\psi = \psi_o \sin(\omega t - kx) \quad \text{and} \quad \psi = \psi_o \sin[(\omega + \Delta\omega)t - (k + \Delta k)x] \quad \dots(8.16)$$

Note that, as we have proven, both these waves are individually faster than light, but if these waves superimpose, their resultant will be

$$\psi = \psi_o \sin(\omega t - kx) + \psi_o \sin[(\omega + \Delta\omega)t - (k + \Delta k)x] \quad \dots(8.17)$$

We know that

$$\sin A + \sin B = 2 \sin\left(\frac{A+B}{2}\right) \cos\left(\frac{A-B}{2}\right)$$

We also know that

$$\cos(-\theta) = \cos(\theta)$$

Applying above identities and rearranging eqn. (8.17) becomes,

$$\psi = 2\psi_o \cos\left(\frac{d\omega}{2}t - \frac{dk}{2}x\right) \sin(wt - kx) \quad \dots(8.18)$$

(We have also used an approximation $2\omega \gg \Delta\omega$ and $2k \gg \Delta k$, as we are superimposing the waves of marginally varying ω and k)

Eqn. (8.18) represents a sine wave whose amplitude is periodically modulated in space and time by a factor $\cos\left(\frac{d\omega}{2}t - \frac{dk}{2}x\right)$. The factor $\cos\left(\frac{d\omega}{2}t - \frac{dk}{2}x\right)$ represents a wave-group which proceeds in space and time with a velocity given by ratio of coefficients of t and x

Thus the group velocity of the wave-group is

$$v_g = \frac{d\omega}{dk} \quad \dots(8.19)$$

We can differentiate, ω w.r.t. k , if ω can be expressed as a function of k

$$E = \frac{p^2}{2m}$$

Also

$$E = h\nu = h\frac{\omega}{2\pi} = \hbar\omega \quad \dots(8.20)$$

$$p = \frac{h}{\lambda} = \frac{h}{2\pi} \times \frac{2\pi}{\lambda} = \hbar k \quad \dots(8.21)$$

Substituting E and p from eqns. (8.20) and (8.21) in the eqn.(8.19), we get

$$\begin{aligned}
 \hbar\omega &= \frac{\hbar^2 k^2}{2m} \\
 \Rightarrow \omega &= \frac{\hbar k^2}{2m} \\
 \Rightarrow \frac{d\omega}{dk} &= \frac{\hbar(2k)}{2m} \\
 \Rightarrow \frac{d\omega}{dk} &= \frac{h}{2\pi} \times \frac{2\pi}{\lambda} \times \frac{1}{m} \\
 \Rightarrow \frac{d\omega}{dk} &= \frac{p}{m} \\
 \Rightarrow v_g &= \frac{d\omega}{dk} = v_{ple} \quad \dots(8.22)
 \end{aligned}$$

Let us recall that the phase velocity of the individual waves which make up the wave-group is still greater than the velocity of light. However, the ultimate velocity of the wave-group which comes into existence after the superposition of individual waves is exactly equal to the velocity of the particle. Thus the wave-group exactly follows the motion of the particle with which it is associated. Thus De Broglie's hypothesis does not violate the theory of relativity



Werner Heisenberg (1901-1976): He studied at University of Munich under Sommerfield and then under Max Born at Göttingen. After receiving his Ph.D. he worked under Max Born and Niels Bohr (both being Nobel laureates) at University of Copenhagen. In 1925, at the age of 23, he published Matrix formulation of Quantum mechanics, which was later shown to be equivalent to its Schrödinger's wave formulation. One of the predictions of his theory was discovery of allotropic forms of hydrogen. For this theory he was awarded a Nobel prize in 1932. Nobel prize was one of the several medals and prizes that he won. His other contributions were plasma physics, thermonuclear processes, and unified theory of elementary particles. What follows is an uncertainty principle known after his name, which he developed at the age of 27.

8.10 HEISENBERG'S UNCERTAINTY PRINCIPLE

Wavelike properties of subatomic particles lead to an unavoidable uncertainty in determining their motion

The quantum mechanics being discussed here is essentially a science of motions of subatomic entities. Let us understand what 'mechanics' means. Mechanics is a discipline which describes the motions of the objects. The parameters which are typically used for describing the motion of an object are position (x), momentum (p), energy (E) and time (t). In classical mechanics, which is based on Newton's second law of motion, the method to obtain these parameters is as follows

For determining the motion of any object (i.e. for determining x, p, E and t), the force acting on the object should be known. This force should be substituted in the equation of motion, i.e. Newton's second law of motion

$$F = ma$$

$$a = \frac{F}{m}$$

$$\frac{dv}{dt} = \frac{F}{m}$$

$$v = \int \frac{F}{m} dt$$

$$p = mv$$

$$E = \frac{p^2}{2m}$$

$$x = \int v dt$$

What is enunciated above is a simplified method of classical mechanics used for determining the motion of an object. If the required mathematics is followed precisely, there seems to be no reason why the quantities like x, p, E and t should not be determinable accurately, precisely (or deterministically). Further, as we know, in our daily lives also, the quantities like velocity (v) and x, p, E and t can be measured with absolute precision with the help of sophisticated instruments. Thus classical mechanics (or our day to day mechanics) is deterministic, and at least principally, it involves no errors or uncertainties.

Can similar approach be followed for subatomic particles? Certainly not! The fundamental reason behind this is that all subatomic particles behave like waves (previously we have seen that because of extremely small value of the Planck's constant, the wavelike properties of day to day objects are negligible). These waves are probability waves, and they are described by ψ . The motions of subatomic particles cannot be analyzed by classical/Newtonian mechanics, as the relation $F = ma$ requires that the particle should behave like a particle and should have definite position in the space. In quantum mechanics, it is ψ which contains entire knowledge about the motions of the subatomic particles. (In next chapter we will learn that the ψ associated with any subatomic particle in any situation can be obtained by solving Schrodinger's equation for that particle). How does ψ contain the knowledge of motion of subatomic particles? Let us look at the simplest form of ψ

$$\psi = \psi_o \sin(wt - kx)$$

$$\psi = \psi_o \sin\left(2\pi vt - \frac{2\pi}{\lambda} x\right)$$

$$\psi = \psi_o \sin\left(\frac{2\pi}{h} hvt - \frac{2\pi}{h} \frac{h}{\lambda} x\right)$$

$$\psi = \psi_o \sin \frac{1}{\hbar} \left(Et - px \right)$$

$$\psi = \psi_o \sin \frac{1}{\hbar} \left(Et - px \right) \dots (8.23)$$

Thus, ψ which represents a De Broglie wave of a given particle, contains the information of all the variables (x , p , E , t) related with the motion of that particle. This can also be understood in a following way.

A wave-group (described using ψ) provides every information required for describing the motion of subatomic particle. The highest peak specifies the most probable position, the distance between consecutive peaks gives λ . $P = \frac{\hbar}{\lambda}$ gives momentum. $E = \frac{p^2}{2m}$ gives energy. But does this wave-group allow us to specify these parameters with as much accuracy as we want? The following discussion shows that the answer is no!

Refer Fig. (8.6). Recall that De Broglie wave is a probability wave. Thus various peaks in $|\psi|^2$ waves represent various probable positions of a particle in various regions. Thus De Broglie wave specifies different probable positions of the same particle in a given region. Does this picture suit our day to day experience? For ex. can we specify different probable positions of a cricket ball on a ground at a given instant? No! Indeed we experience that a cricket ball existing on a ground at a given instant has only one definite position (and therefore only it can be picked up!).

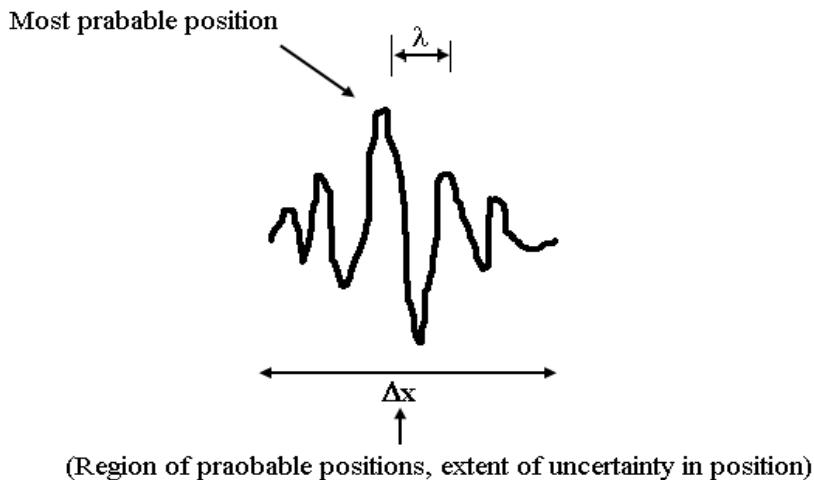


Figure (8.6): A wave-group associated with a particle contains entire information related with the motion of the particle

Let us also recall that the De Broglie waves associated with cricket ball (and every day to day object) are too feeble to be considered. The very fact that a De Broglie wave specifies

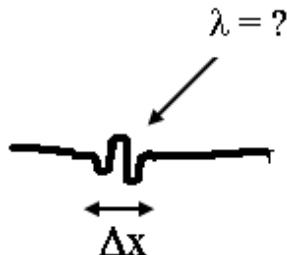
different probable positions of the same particle at a given instant clearly indicates that the association of De Broglie waves imposes an unavoidable uncertainty in determining the position of a particle. The size of De Broglie wave indicates the region of uncertainty in the position and therefore it can be taken as uncertainty/error “ Δx ” in determining the position of the particle. Now, let us pay attention to another equally important parameter i.e. momentum “ p ” of the particle. Can it be specified deterministically? No! Let us see how.

According to De Broglie's hypothesis

$$p = \frac{h}{\lambda}$$

Thus deterministic description of p requires deterministic description of λ ; and a glance at the De Broglie wave-group indicates that De Broglie wave-group does not have a precise wavelength. The distance between various neighboring peaks varies from peak to peak. This means that the association of De Broglie wave with a particle also imposes an unavoidable uncertainty in describing its wavelength and hence the momentum.

Now let us see whether the uncertainty in position (Δx) and uncertainty in momentum (Δp) can be minimized. (in Physics, we always aim at minimizing the errors). We will now see that these can be minimized, but not simultaneously (and this is the fundamental concept of Heisenberg's uncertainty principle.). Let us see how.



Tiny wavegroup due to superposition
of extremely large number of waves
 Δx very small but $\Delta \lambda$ hence Δp is very
large

Figure (8.7): Tiny wave-group

Consider an extremely compact wave-group shown in the Fig. (8.7). Such wave-group can be obtained by superposition of an extremely large number of waves, each having different wavelengths. (If the number of interfering waves, each having different wavelengths is infinity, then the size of resultant wave-group will become zero. In our case, the number of interfering waves is almost infinity, but not exactly infinity, so the size of resultant wave-group tends to zero, but does not become absolutely zero...refer Fig (8.7 once again). Thus Δx which is of the order of size of the wave-group becomes extremely small (but not zero!). But, at what cost? As

it can be noticed from the Fig (8.7), the wave-group is so compact that it is extremely difficult to measure its wavelength. Indeed, smaller the wave-group, more undefined is its wavelength; and as wavelength becomes uncertain, the momentum also becomes uncertain. Also remember that such tiny wave-group has been formed by the interference of many waves, each having different wavelength. Thus fundamentally the waves in such groups have very large variations in the wavelengths. Thus the waves involved in this case have no definite wavelength. Thus when Δx becomes smaller and smaller and smaller, Δp becomes larger, and larger. Simultaneous accuracy in position (x) and momentum (p) is not possible in this case.

Now let us consider the case as depicted in the following Fig (8.8). In case of wave-group of the medium size, both Δx and $\Delta \lambda$ and hence Δp are fairly moderate, but none of them tends to zero. Thus simultaneous accuracy in x and p is not possible in this case also.

Now let us start approaching to another end of the situation. Refer Fig (8.9). We will start widening the wave-group (as we want to make momentum and hence the wavelength more accurate). In doing so, we will have to decrease the number of interfering waves. Looking at the size and shape of resulting widened wave-group we find it fairly easy to find out the wavelength. Observe that there are sufficient numbers of peaks available to establish the wavelength now. Further, due to decrease in the number of interfering waves having different wavelength, the ‘spread’ in $\Delta \lambda$ is relatively small, thus Δp also decreases. But note that in doing so we have widened the wave-group. Thus the particle which is supposed to have one definite position at a given instant has many probable positions spread in the widened wave-packet. The conclusion is that Δx has increased.

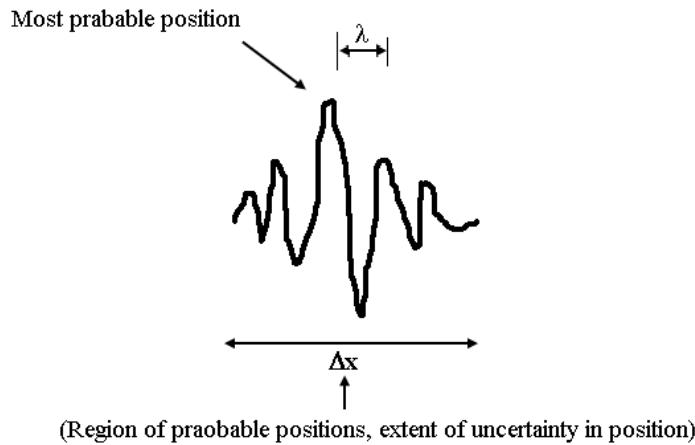


Figure (8.8): A wave-group of medium size. Here Δx as well as $\Delta \lambda$ hence Δp are moderate. But none of them tends to zero

**λ is more relatively more accurate.
 $\Delta\lambda$ and hence Δp decreases**

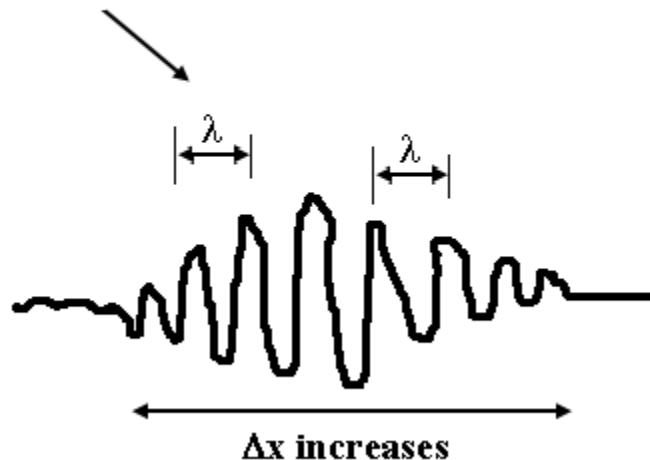


Figure (8.9): Relatively wide wave-group

An infinitely long wavetrain
(wavegroup containing single wave).
 λ is same everywhere. Thus $\Delta\lambda$ i.e. Δp is zero

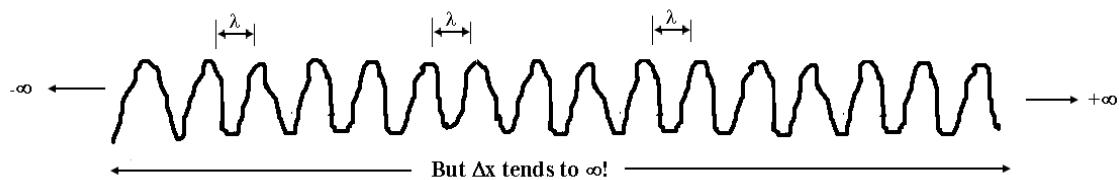


Figure (8.10): Infinitely wide wave-group

An extreme end that can be reached by widening the wave-group is to obtain single but infinitely long wave (Refer Fig. 8.10). The number of interfering waves has reduced to one. Thus, we may consider any two neighboring peaks, the distance between them, i.e. wavelength is same. The wavelength being same everywhere, the ‘spread’ in λ , i.e. $\Delta\lambda$ is absolutely zero. Thus in such extreme case, we have succeeded in making $\Delta\lambda$ i.e. Δp zero. But, at what cost? We know that a single wave (or a wave-group containing a single wave) is always infinitely long. Thus Δx , which is related to size of the wave-group has become infinite!. Our attempts to achieve simultaneous accuracy in x and p have once again failed! We thus conclude that it is not possible to construct a wave-group of any size or shape, which gives us simultaneous accuracy in x and p . This is what Heisenberg’s uncertainty principle means.

Now let us see whether we can express our discussion in quantitative terms. It can be shown that the size of wavegroup is $\frac{\lambda_{av}^2}{\Delta\lambda}$, where λ_{av} is the average wavelength of all the waves and $\Delta\lambda$ is the spread in their wavelengths

$$\text{Thus } \Delta x = \frac{\lambda_{av}^2}{\Delta \lambda}$$

We have

$$\begin{aligned} P &= \frac{h}{\lambda} \\ P &= h\lambda^{-1} \\ \Delta p &= -h\lambda^{-2}\Delta\lambda \end{aligned}$$

If Δp is considered as an error in p , then

$$\begin{aligned} \Delta p &= |-h\lambda^{-2}\Delta\lambda| \\ \Delta p &= \frac{h}{\lambda^2} \Delta\lambda \end{aligned}$$

Taking the product of Δx and Δp , we get

$$\Delta x \Delta p = h$$

Thus the minimum value of the product of Δx and Δp is h . The product cannot be less than h . Further, as error has no upper limit

$$\Delta x \Delta p \geq h$$

A more accurate approach give

$$\Delta x \Delta p \geq \frac{\hbar}{2} \quad \dots(8.24)$$

Thus the Heisenberg's uncertainty principle states that in any attempt of determination of position and momentum, simultaneous accuracy in both is not possible. The product of uncertainty in position and uncertainty in momentum can never be reduced below h

8.11 AN EXPERIMENTAL PROOF OF UNCERTAINTY PRINCIPLE I: SINGLE SLIT ELECTRON DIFFRACTION

The width of slit affects the uncertainties in position and momentum in an opposite manner

Refer Fig. (8.11). Consider a beam of electrons passing through a slit having width comparable with the wavelength of electrons. Electrons, being waves, will get diffracted through the slit and form a diffraction pattern on the slit.

Now let us focus our attention on one out of several electrons passing through the slit. This situation essentially corresponds to a thought experiment. There are two reasons behind considering this as a ‘thought experiment’. One is that it is never possible to construct a single slit having its dimensions comparable with the wavelength of electrons. Secondly it is just impossible to focus our attention one out of several electrons getting diffracted through the slit.

Now let us ask two questions and try to answer them

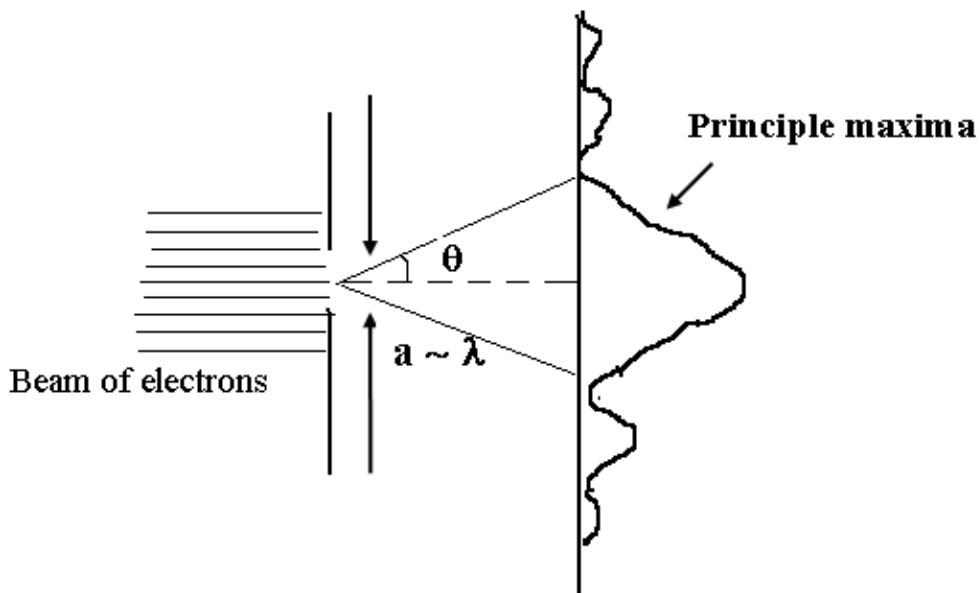


Figure (8.11): Single slit diffraction pattern of electrons

- i. From which part of the slit, one out of several electrons is passing?
- ii. What is the exact value of Y component of the momentum of the electron when it is diffracted through the slit?

The nature of these questions apparently indicates that uncertainties should be involved in describing the position and momentum of the electron while passing through the slit.

Though several electrons, after passing through the slit, produce a well defined diffraction pattern on the slit, predicting the exact position of one electron when it passed through the slit is extremely difficult as it can pass through any part of the slit. Thus uncertainty involved in determining the position of the electron while it is passing through the slit is

$$\Delta y = a, \text{ where } a \text{ is the width of the slit}$$

From the theory of single slit diffraction, the minima is described by

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

Thus, $\Delta y \sin \theta = n\lambda$

We need to take $n = 1$, as the principle image of slit is mainly reflected in the principle (central) maxima and the minima associated with the central maxima is the first minima . Secondary maxima are too weak to be considered. Thus

$$\Delta y \sin \theta = \lambda$$

$$\Delta y = \frac{\lambda}{\sin \theta} \quad \dots(8.25)$$

Now let us pay attention to momentum. The electron doesn't possess momentum in y direction before passing through the slit, its entire momentum is in x direction. According to De Broglie's hypothesis

$$p = \frac{h}{\lambda}$$

While passing through the slit, electron takes part in the diffraction process. Therefore, electron acquires a y component of momentum after getting diffracted. Note that the well defined principle maximum is the result of diffraction of several electrons. Out of such several electrons, the y component of a single electron is almost unpredictable. After diffraction, a single electron possesses x as well as y components of momentum, which are given by $\frac{h}{\lambda} \sin \theta$ and $\pm \frac{h}{\lambda} \cos \theta$ respectively. As the electron while approaching the principle maxima has a freedom to proceed at any angle within a 'cone having angle 2θ ', it may acquire a y component of momentum which may have any value from $\frac{h}{\lambda} \sin \theta$ to $-\frac{h}{\lambda} \sin \theta$ where θ is the angle of diffraction of the first minimum. (Note that in this discussion, we are always considering the principle maximum, which is bounded on both the sides by first minimum. The secondary maxima are too weak to be considered). Thus, uncertainty in y momentum is

$$\Delta p_y = 2 \frac{h}{\lambda} \sin \theta \quad \dots(8.26)$$

Taking the product of Δy and Δp_y from Eqns. (8.25) and (8.26), we get

$$\Delta y \Delta p_y = 2h$$

The minimum value of the product is

$$\Delta y \Delta p_y = h$$

And as error has no upper limit,

$$\Delta y \Delta p_y \geq h \quad \dots(8.27)$$

This indicates that as Δy increases Δp_y decreases, and vice versa. Simultaneous accuracy in position and momentum is thus not possible. It can also be observed that a smaller slit will yield less uncertainty in position, but then, as diffraction occurs more strongly for a tiny slit, the θ will be large leading to higher uncertainty in momentum. And wide slit will increase the uncertainty in position, but will decrease the uncertainty in momentum, as for wider slit, the diffraction and hence θ will be small. There doesn't exist a slit which can give simultaneous and minimum uncertainty in position and momentum at the same time. This itself is the crux of Heisenberg's uncertainty principle.

8.12 ANOTHER APPROACH AT HEISENBERG'S UNCERTAINTY PRINCIPLE

A subatomic particle cannot be observed without disturbing it

For determining any parameter related with the motion of particle x, p, E and t, any subatomic particle, say an electron, needs to be 'seen', and therefore needs to be illuminated. For illumination we need light. During illumination the photons in the light will fall on the electron and electron being one of the lightest particles, its position as well as momentum will be disturbed. To minimize the disturbance (and hence the uncertainty), we minimize number of photons to only one. This photon will collide with the electron, will get scattered, and then will enter in the microscope and its interaction with the microscope will provide us the knowledge of position and momentum of the electron. The collision between photon and electron is thus unavoidable. But the collision process itself will disturb the original position and momentum of electron in an unpredictable manner. Thus uncertainties are unavoidable in this case also. The crux of this situation is that the photon cannot provide the information about the position and momentum of electron without disturbing them in an unpredictable manner. Now let us choose the photon and a microscope. Gamma ray photon and Gamma ray microscope is the best choice, because gamma rays have the lowest wavelength and lower the wavelength, higher is the resolution. For determining the position and momentum of the electron, the gamma ray photon has to fall on the electron; disturb the position and momentum of electron in an unpredictable manner and will then enter in to the microscope thorough any point of its aperture. Now let us consider two kinds of microscopes. One, a gamma ray microscope of wider aperture using gamma ray photon of smaller wavelength and another, a gamma ray microscope of small aperture, using gamma ray photon of relatively larger wavelength. In first case the measurement of position will be more accurate as the resolving power of gamma ray microscope increases with increase in the size of aperture and with decrease in wavelength. However in this case uncertainty in the momentum will increase as the photon of smaller wavelength will carry higher momentum and will hit upon the electron strongly and will cause more disturbance in momentum. Further, after colliding with the electron, this photon will have a freedom to enter in the microscope through a wider cone, as the aperture of microscope is large. This will also create greater difficulty in determining the original momentum of electron. We can make exactly opposite arguments and show that a microscope with narrow aperture using a photon of larger

wavelength will measure momentum with more accuracy but position with less accuracy. Thus it is impossible to choose a microscope which will give simultaneous accuracy in position as well as momentum. Just like the single slit electron diffraction experiment, this is also a thought experiment, as a gamma ray microscope does not exist and it is impossible to create a situation where only one gamma ray photon can be made to fall on one electron.

Several such thought experiments were proposed to prove and to disprove this principle. The thought experiments for proving the principle were proposed by a group of physicists including Niels Bohr, Heisenberg etc., while the experiments attempted for disproving the principle were proposed by Albert Einstein, who did not accept this principle till his death. Einstein never accepted the fact that there can be an inherent uncertainty in the physical variables, as a law of nature itself. His argument was ... "**God does not play dice**". But, our previous discussion indicates that it is not possible to toss around this principle. Atleast in this case, nature deceived the genius!

There is one more important aspect of Heisenberg's uncertainty principle. In mechanics, to predict the future of a moving particle, it's position and momentum at 'present' must be known with absolute certainty. We have seen that this is impossible. Thus future of subatomic particles also is unpredictable. The subatomic world is unpredictable

Heisenberg's uncertainty principle in terms of energy and time:

We know that

$$E = \frac{1}{2}mv^2$$

$$\Delta E = \frac{1}{2}m2v\Delta v$$

$$\Delta E = (m\Delta v)v$$

$$\Delta E = \Delta p \frac{\Delta x}{\Delta t}$$

$$\Delta E \Delta t = \Delta p \Delta x \geq h$$

$$\Delta E \Delta t \geq h$$

In more accurate terms

$$\Delta E \Delta t \geq \frac{\hbar}{2} \quad \dots(8.28)$$

8.13 HEISENBERG'S UNCERTAINTY PRINCIPLE AND QUANTUM MECHANICS

Heisenberg's uncertainty principle is consistent with several predictions of quantum mechanics

Let us consider minimum value of the product of ΔE and Δt

$$\Delta E \Delta t = h$$

Ground state energy is defined as the minimum possible energy that a subatomic particle can have. According to our day to day standards, this energy can be zero (a cricket ball can remain at rest on the cricket ground). But this is not possible for subatomic particles. (we have already proved this using De Broglie's hypothesis). Let us see how.

If $E = 0$, then $\Delta E = 0$, this will yield $\Delta t = \infty$. This will give $t = \infty$. This will mean that the particle will rest in zero energy state forever; it will never be possible to raise its energy. But we know that electrons can be excited to higher energy states. This means that ground state energy of electron can never be decreased to zero. This means that electron and any subatomic particle can never take a rest. The world at atomic level is restless!

This conclusion can also be arrived at by using $\Delta x \Delta p = h$. If we let $E = 0$, then $p = 0$. This will give $\Delta p = 0$ and thus $\Delta x = \infty$. Thus if we let electron take a rest, then uncertainty in finding its position will reach infinity. This is not possible as electron always exists in confined spaces such as atom, molecule or CRT etc. The uncertainty in finding the position of the electron existing in confined spaces having finite boundaries can never reach to infinity. The conclusion is that according to quantum laws, electron, and any subatomic particle can never brought to rest.

Now let us discuss quantization in the light of Heisenberg's uncertainty principle. As we shall learn in detail in the next chapter, the energy of an electron in an atom is quantized. This means that electron can exist only in discrete energy levels. All energy levels are not possible, some are forbidden. In another language the difference between any two energy levels of electrons has to be finite. This is also consistent with $\Delta E \Delta t = h$. If we assume a continuous and non-discrete energy spectrum of a subatomic particle, then the energy levels will be very closely spaced. This means that the difference between consecutive energy levels will tend to zero i.e. $\Delta E \rightarrow 0$, but this will result in $\Delta t \rightarrow \infty$. This means that if we try to excite or de-excite an electron it will take infinite time to do so. In another language, it will not be possible to excite or de-excite an electron. But we know that electrons can be excited or de-excited. For this to be possible the energy spectrum of electron/any subatomic particle has to be discrete. Quantization of energy is inherent in the world of subatomic particles.

The Heisenberg's uncertainty principle brings one more fact of the nature before us. $\Delta E \Delta t = h$ indicates that ΔE cannot become zero as Δt cannot become ∞ . We know that $E = hv$. Thus $\Delta E = h \Delta v$. If ΔE is nonzero, then Δv is also nonzero. This also indicates that spectral lines cannot have sharp frequency (and thus the wavelength).

Now let us discuss a few more aspects of Heisenberg's uncertainty principle by solving some problems.

Example (8.10):

Use Heisenberg's uncertainty principle to prove that electron cannot exist inside the nucleus (Size of nucleus $\approx 10^{-14}$ m)

Solution

We know that $\Delta x \Delta p = h$. As the size of nucleus is approximately 10^{-14} , if electron exist inside the nucleus, then

$$\Delta x \sim 10^{-14} \text{ m}$$

$$\Delta x \Delta p = h$$

$$\Delta p = \frac{h}{\Delta x}$$

$$\Delta p = \frac{6.63 \times 10^{-34}}{10^{-14}}$$

$$\Delta p = 6.63 \times 10^{-20}$$

$$m \Delta v = 6.63 \times 10^{-20}$$

Thus

$$\Delta v = \frac{6.63 \times 10^{-20}}{9.1 \times 10^{-31}}$$

$$\Delta v = 7.29 \times 10^{10} \text{ m/s}$$

As the physical quantity always greater than its error

$$v \geq 7.29 \times 10^{10} \text{ m/s}$$

Thus if electron is allowed to exist inside the nucleus, its speed will exceed the speed of light. This violates the principle of special theory of relativity. Thus electron cannot exist inside the nucleus.

As an another approach

$$\Delta p = 6.63 \times 10^{-20} \text{ kg.m/s}$$

The smallest value of a physical quantity can be error in itself. Thus

$$p_{\min} = 6.63 \times 10^{-20} \text{ kg m/s}$$

$$E_{\min} = \frac{p_{\min}^2}{2m}$$

$$E_{\min} = \frac{(6.63 \times 10^{-20})^2}{2 \times 9.1 \times 10^{-31}}$$

$$E_{\min} = 2.42 \times 10^{-9} \text{ J}$$

$$E_{\min} = \frac{2.42 \times 10^{-9} \text{ J}}{1.6 \times 10^{-19} \text{ J/eV}}$$

$$E_{\min} = 1.5 \times 10^{10} \text{ eV}$$

$$E_{\min} = 15095 \text{ MeV}$$

Thus if electron existed inside the nucleus then its minimum energy would have been 15095 MeV, which is far greater than the maximum binding energy of the nucleus (8.8 MeV). Thus Heisenberg's uncertainty principle is consistent with the fact that electron does not exist inside the nucleus. (Then what about β rays? Do they exist inside the nucleus? No! β rays are emitted away from the nucleus when a neutron is converted into proton).

Non-existance of the electron inside the nucleus and the discovery of neutron

After the discovery of proton, there was a realization that the number of protons alone could not explain the mass of the nucleus. This is because, it was known that the atomic mass is roughly double the atomic number (atomic number is number of electrons in the atom, and consequently the number of protons in side the nucleus). As almost entire mass of the atom is concentrated inside the nucleus, nucleus would require, as many 'proton-like' but electrically neutral partciles as the number of protons. It was at first thought that such additional partcile could be proton itself, but surrounded by an electron, so that the net charge of the proton-electron pair would be neutral. But quantum mechanics decisively proved that electron can not exist inside the nucleus. Thus instead of proton-electron pair, an independant particle, as heavy as proton, but electrically neutral would solve the puzzle. Rutherford postulated this partcile as neutron, but the experimental confirmation of the neutron, which was an experimentally difficult task, was done by James Chadwick (as student of Rutherford) who discovered the neutron. The use of neutron in nuclear fission gave birth to the nuclear energy. James Chadwick was awarded a Nobel prize in Physics in 1935 for discovering neutron.

Example (8.11)

An electron is orbiting around the nucleus with a velocity $2 \times 10^5 \text{ m/s}$. The uncertainty in its velocity is 5%. How much is the corresponding Heisenberg's uncertainty in its position? A cricket ball is moving at the velocity of 20 m/s, with an uncertainty of 5%. How much is the corresponding uncertainty in position? The mass of cricket ball is 0.5 kg.

Solution:

For electron

$$\Delta v = \frac{5 \times 2 \times 10^5}{100}$$

$$\Delta v = 10^3 \text{ m/s}$$

$$\Delta x \Delta p = h$$

$$\Delta x m \Delta v = h$$

$$\Delta x = \frac{h}{m \Delta v}$$

$$\Delta x = \frac{6.63 \times 10^{-34}}{9.1 \times 10^{-31} \times 1000}$$

$$\Delta x = 7.28 \times 10^{-7} \text{ m}$$

$$\Delta x = 7286 \text{ A}^\circ$$

This error is quite considerable looking at the fact that radii of atomic orbits are in a few A°

$$\text{For Cricket ball } \Delta v = \frac{5 \times 10}{100}$$

$$\Delta v = 1 \text{ m/s}$$

$$\Delta x \Delta p = h$$

$$\Delta x = \frac{h}{m \Delta v}$$

$$\Delta x = \frac{6.63 \times 10^{-34}}{0.5 \times 1}$$

$$\Delta x = 1.326 \times 10^{-33} \text{ m}$$

As it can be noticed, for the same 5% uncertainty in the velocity of electron and cricket ball, the uncertainty in the position of electron is considerable and uncertainty in the position of cricket ball is negligible. This indicates that Heisenberg's uncertainty principle (and as we shall learn, every principle in quantum mechanics) works only in world of subatomic particles. The uncertainties indicated by this principle are too small to be considerable for the day to day objects. This is due to the typical (an extremely small) value of Planck's constant. What would happen, if Planck's constant possessed a different value? Let us solve a problem.

Example (8.12)

A cricket ball is moving at 20 m/s with 5% uncertainty. What would be the corresponding uncertainty in the position, if Planck's constant were 6.63 J-S? Assume mass of cricket ball to be 0.5 kg.

Solution

$$\Delta v = \frac{5 \times 20}{100} = 1 \text{ m/s}$$

$$\Delta x \Delta p = h$$

$$\Delta x = \frac{h}{m \Delta v}$$

$$\Delta x = \frac{6.63}{0.5 \times 1} = 1326 \text{ m}$$

This uncertainty is too large to be neglected. It would be extremely difficult to catch the cricket ball, if Planck's had such value. This would happen to every object, if Planck's constant had this value. One can notice that nature has cleverly chosen the value of Planck's constant to be 6.63×10^{-34} J.s, to make us comfortable!

As a conclusion of problems solved in this chapter, we may state that classical mechanics is the science of motion of day to day objects such as cricket ball, projectiles, airplanes, satellites, planets etc., while quantum mechanics is the science of motion of subatomic particles such as electrons, protons, neutrons, atoms and molecules. And it is the Planck's constant, which differentiates classical mechanics from quantum mechanics. The reason that quantum mechanics and classical mechanics work in their own realms is because the value of Planck's constant is 6.63×10^{-34} . Had it been different, probably our day to day motions would also be determined by quantum laws

8.14: PARTICLE IN A RIGID BOX: WITH FIRST PRINCIPLES

Restriction leads to quantization

We now aim to apply the basic principles which we have learnt to a typical problem in quantum mechanics named 'particle in rigid box'. This will help us to take in to account a few characteristic features of quantum mechanics. The problem 'particle in a rigid box' has many

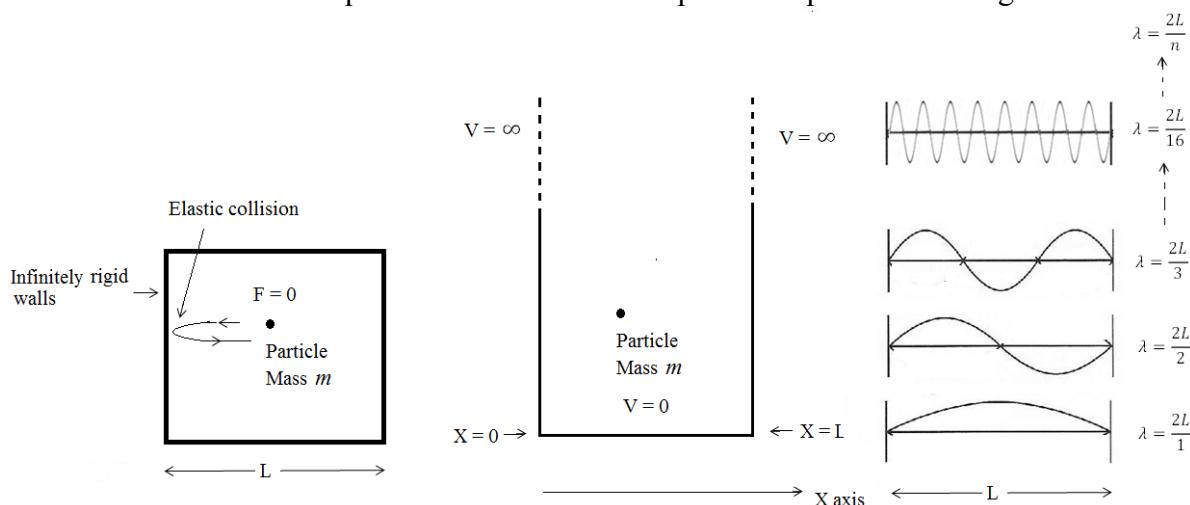


Figure (8.12): Particle in a rigid box: Problem and solution

analogies in day to day life, science and even technology. Indeed all motions in the nature are in a way, the motions in boxes. For ex. motions of an electron in the atom, proton or neutron in nucleus, atom in a molecule are all the examples of motions of the particles in certain kinds of ‘boxes’. Even, the motion of a cricket ball on the ground is also a kind of motion in a ‘box’. Here, we treat atom, nucleus, molecule and even a cricket ground as a ‘box’. The fundamental feature of all these motions is that all of them are the motions with restrictions. The particles that we consider here are not the free particles. They are bounded within the boxes. Science and technology is rapidly running towards nano-age. The devices, machines and materials are becoming so tiny that, they have their sizes in the range of few nanometers, and they contain limited number of atoms. The motion of electron in such tiny nanosystems (for ex. quantum dot, or a quantum corral or a nanocluster) can also be considered as a motion in a box.

The boxes that we have mentioned here are not simple ones. The particles in these boxes can move in all three dimensions. Moreover they move in complex environments. As we shall notice it further, the analysis of these motions require quite elaborate mathematics. Therefore we simplify our discussion by making certain assumptions, which are

- i. The particle moves only in one direction , say X axis
- ii. The force acting on the particle is zero. Thus, though the particle is trapped in the box, it is free within the box. The potential energy of the particle inside the box is thus zero, and consequently its entire energy is kinetic energy
- iii. The walls of the box are infinitely rigid. Thus the collisions of the particle with the walls are elastic and there no loss of energy during the collisions. The energy of particle is thus constant
- iv. Another consequence of assuming infinitely rigid box is that, the particle with whatsoever energy can never come out of the box. Thus we need to solve the problem only inside the box.

Such boxes can be considered as one dimensional trap. We recall that in quantum mechanics, we treat the particle as a wave. With the simplifications mentioned above, we can perfectly correlate this problem with the standing waves generated on a string tied with rigid supports at both the ends.

Recall that we can generate a travelling wave of any wavelength; small, medium or large, on a string of infinite length. There are no boundaries and thus no boundary conditions. However, once we tie the string between the rigid supports, we expect it quite apparently that the string should not oscillate at the rigid boundaries. Thus only those standing waves can be generated, which will have nodes...the points of zero displacements, at the rigid boundaries. We very well know that such standing waves satisfy a fundamental equation given by

$$\lambda = \frac{2L}{n} \quad \dots (8.29)$$

The condition of zero displacement at the rigid boundaries requires that n be an integer. Further, we cannot start with $n = 0$, because, then the equation gives λ to be infinity. We cannot generate a wave of infinite wavelength on a string of finite length. Thus we have $n = 1, 2, 3, \text{ etc.}$, and corresponding wavelengths are $\lambda = \frac{2L}{1}, \frac{2L}{2}, \frac{2L}{3}, \dots, \frac{2L}{16}, \dots$ etc. Note that for any fractional value of

n , the standing waves will have non-zero displacements on the boundaries. The boundaries being rigid, this is not permissible. The n is thus restricted to have only integer value (except zero). In the language of quantum mechanics n is called “*quantum number*” (quantization means restriction). If the waves in Fig (8.12) are the matter waves, we have

$$\lambda = \frac{h}{p}$$

This gives

$$p = \frac{h}{\lambda} = \frac{h}{\frac{2L}{n}} = \frac{nh}{2L}, \quad \text{with } n = 1, 2, 3, 4 \dots \quad \dots(8.30)$$

Momentum of a particle in a rigid box is thus quantized. Also note that De Broglie’s hypothesis as well as Heisenberg’s uncertainty principle do not permit exactly zero momentum, which gives an additional justification for only nonzero allowed values of n .

We recall that in our case, entire energy of the particle is kinetic energy. Thus,

$$E = KE = \frac{1}{2}mv^2 = \frac{p^2}{2m} = \frac{\left(\frac{nh}{2L}\right)^2}{2m}$$

$$\Rightarrow E = \frac{n^2 h^2}{8mL^2}, \quad \text{where } n = 1, 2, 3, 4 \text{ etc} \quad \dots(8.31)$$

As n is not permitted to have zero value, $E = 0$ is not permissible. The minimum energy that the particle can have is thus

$$E_1 = \frac{h^2}{8mL^2} \quad \dots(8.32)$$

This energy is called as a ‘*ground state energy*’ or ‘*zero point energy*’. The next few permissible energies are $\frac{h^2}{8mL^2}, \frac{4h^2}{8mL^2}, \frac{9h^2}{8mL^2}$ etc. Intermediate energies require fractional values of n and thus are not permitted. Thus the energy of a particle in a rigid box is quantized.

We can easily trace the origin of quantization in this problem. We have restricted the motion of a particle within rigid boundaries. The wave nature of particle requires that such particle be represented by only standing waves having points of zero displacements (nodes) at the boundaries. This requires $\lambda = \frac{2L}{n}$, which leads to eqns (8.30) and (8.31). Recall that such restrictions are not applicable to the waves on infinitely long string, where a travelling wave of any wavelength is possible. This is equivalent to an entirely free particle, which can move anywhere in infinite space without any restriction. The quantization is a direct consequence of the fact that the particle is bounded. We thus conclude that

(i). A free particle is permitted to have any energy but the energy of a bounded particle is always quantized. We recall that all motions in the universe are the motions in ‘boxes’. Indeed a ‘free particle’ is entirely an idealization. All entities in the nature, may it be an electron in an atom or a cricket ball on a ground, are bounded i.e. their motions are restricted. We thus conclude that *restriction on the motion leads to quantization*.

It is quite necessary here to differentiate once again between the objects in daily life (say, a cricket ball on a ground) and those in the subatomic world (say an electron in an atom). For this purpose, we write eqn (8.31) for two discrete quantum numbers, say p and q . Thus we have

$$E_p = \frac{p^2 h^2}{8mL^2} \quad \text{and} \quad E_q = \frac{q^2 h^2}{8mL^2}$$

Subtracting,

$$E_p - E_q = \frac{h^2}{8mL^2} (p^2 - q^2)$$

Recall that, $h^2 \approx 10^{-68}$. For the subatomic particles moving in subatomic regimes, the mass m and the length of ‘box’ are considerably small (for ex the mass of electron is $9.1 \times 10^{-31} \text{ kg}$ and the length of ‘atomic’ box is roughly 1 \AA giving $L^2 \approx 10^{-20}$). The numerator and the denominator in the term outside the bracket are thus equally small. Thus the ratio is significant. Thus the energy levels E_p and E_q for any two discrete quantum numbers p and q are sufficiently away. This means that the energy levels of a subatomic particle trapped in a tiny zone are discrete i.e. quantized.

Does this apply to our daily life? Certainly (and fortunately) not!. For ex. consider a ball of mass 1 kg moving in a box of length 1 m . In such case the numerator in eqn (8.31) comes out to be extremely small as compared to the denominator. Thus for any two arbitrary values of p and q , the energy levels E_p and E_q are extremely close to each other, in fact so close that the energy spectrum can be considered to be continuous. This discussion leads to an important conclusion that the principle of quantization works only in subatomic world and not in our daily life.

(ii). We have also noted that the ground state energy of a particle in a rigid box is given by

$$E_1 = \frac{h^2}{8mL^2}$$

We once again recall that for subatomic particles, the denominator in above eqn. is almost as small as the numerator, leading to a finite, significant and non-zero value of E_1 . Thus the ground state energy of a particle is non-zero. This also means that a state with exactly zero energy is forbidden. The subatomic world is thus a restless world. The subatomic particles cannot take rest (they can be made slow but cannot be brought to rest). However, for the objects in daily life, both m and L and hence the denominator in above eqn are considerably large as compared to numerator. Thus the ground state energy of any object in our daily life is too small to be considerable. Thus objects in our daily life can thus be brought to rest.

In both the principles discussed above, the quantization and non-zero ground state energy is off course due to extremely small value of the Planck's constant. Imagine that we are living in a universe where Planck's constant is simply $\hbar = 6.63 \text{ J.s}$. Our daily life will then be quantized. We will be allowed to walk or run with only certain allowed speeds, and the players will be allowed to hit the ball only in certain ways and so on! Further, our ground state energies will be significant. We will thus not be allowed to be at the rest! Wait!, Don't be anxious, the 'nature' is clever enough to choose a right value of Planck's constant so that we live comfortably in the normal classical world!

The expression for ground state energy of a particle in rigid box can also be obtained by using Heisenberg's uncertainty principle. For a particle trapped in a box of length L (if sufficiently small), the error in locating its position can be roughly taken to be

$$\Delta x \sim 2L$$

The factor '2' accounts for the fact that particle can move in $\pm x$ direction. The Heisenberg's uncertainty principle is

$$\Delta x \Delta p \approx h$$

Substituting Δx , we get

$$2L \times \Delta p \approx h$$

$$\Rightarrow \Delta p \approx \frac{h}{2L}$$

In subatomic world, the momenta are extremely small. Thus

$$p \approx \Delta p \approx \frac{h}{2L}$$

Energy is

$$E = \frac{p^2}{2m} \approx \frac{h^2}{8mL^2}$$

SUMMARY

As radiations exhibit dual character, the matter can also have dual character. A wave called as De Broglie wave is associated with every moving particle. These waves have a wavelength given by $\lambda = h/p$. This is called as De Broglie's hypothesis. These waves are probability waves. Being

wave-groups, their energy is quantized. The De Broglie wave-group moves with the same speed of the particle with which it is associated.

The wavelike properties of subatomic particles impose an unavoidable uncertainty in determining their position, momentum, energy and time. This is called as Heisenberg's uncertainty principle. Mathematically

$$\Delta x \Delta p \geq \frac{\hbar}{2} \text{ and } \Delta E \Delta t \geq \frac{\hbar}{2}$$

The above mathematical statements indicate that simultaneous and unlimited accuracy in position and momentum or energy and time are not possible.

Both De Broglie's hypothesis and Heisenberg's uncertainty principle point out towards two fundamental principles in quantum mechanics. These are nonzero ground state energy and quantization of the energy.

EXERCISES

Questions are never indiscreet, answers sometimes are.

Oscar Wilde

1. Explain in your own words and briefly the significance of Quantum mechanics in engineering and technology
2. What made De Broglie to make a suggestion that a material particle behaves like a wave?
3. Does De Broglie's hypothesis have sound experimental foundation? What is it?
4. Give any three examples which make Physicist's believe that "Nature loves symmetry" Can you give at least one example of your own in addition to those given in this book?
5. Consider following formulae

$$\lambda = \frac{h}{p} \text{ and } E = hv$$

In the first relation λ on L.H.S. is a wave property and p on R.H.S. is a particle property. In second relation also E on L.H.S. is a particle property and v on R.H.S. is a wave property. How the same equation can have both particle property and wave property existing together?

6. At some place in this book we have equated two formulae for the energy of an electron...these are $E = hv = mc^2$. How an equation can equate the formulae for energy where one formula considers electron as a particle and another considers electron as a wave?
7. Give at least one similarity between a photon and material particle and at least one difference between them.
8. Explain how De Broglie's hypothesis is consistent with Bohr's quantization of electron orbits.
9. "*De Broglie waves are probability waves*". Justify

10. “*De Broglie waves are wave-groups*” Justify
11. “*De Broglie waves are quantized*” justify
12. Compare De Broglie waves with sound waves in as many respects as you can.
13. Compare De Broglie waves with light waves in as many respects as you can.
14. De Broglie’s hypothesis works conspicuously for subatomic particles but not for day to day objects. Why?
15. According to De Broglie’s hypothesis, subatomic particles can’t take a rest. Why?
16. Though, principally the De Broglie’s hypothesis does not permit any entity to take rest, the objects in our day to day life are observed to be taking rest. How?
17. What do you exactly mean by ground state energy?
18. What do you exactly mean by quantization of energy?
19. Differentiate between wave, particle and wavicle.
20. According to De Broglie’s hypothesis, the energies of the subatomic particles are quantized but the energies of the day to day objects are not quantized. Why?
21. In quantum mechanics it is not ψ , but it is $|\psi|^2$ which is physically interpreted as probability density. Why?
22. Why $|\psi|^2$ is interpreted as probability density and not simply as probability?
23. “*De Broglie wave-group has two velocities, one is the phase velocity and another is the group velocity*”. Comment
24. “A free particle is associated with an infinitely long single progressive wave, but a confined/bounded particle is always associated with a wave-group”. Comment.
25. What do you exactly mean by wave-group? How it is formed? Comment on narrow wave-group and wide wave-group.
26. Comment on “*Energy of a free particle is not quantized but energy of a bounded particle is quantized*”
27. Comment on “***Everything in future is wave, everything in past is particle***”
28. What is Heisenberg’s uncertainty principle?
29. Which physical quantity out of position and momentum is more certain in the narrow wave-group? Why?
30. Which physical quantity out of position and momentum is more certain in a wide wave-group? Why?
31. Under what conditions a narrow wave-group is formed? Under what conditions a wide wave-group is formed?
32. In single slit diffraction, which quantity becomes certain on widening the slit? Why? Which quantity becomes certain on narrowing the slit? Why?
33. Comment on “*There doesn’t exist a microscope which will give simultaneous accuracy in position as well as momentum*”
34. How De Broglie’s hypothesis is consistent with nonzero ground state energy?
35. How De Broglie’s hypothesis is consistent with quantization of energy?
36. How Heisenberg’s uncertainty principle is consistent with nonzero ground state energy?
37. How Heisenberg’s uncertainty principle is consistent with quantization of energy?
38. According to Heisenberg’s uncertainty principle, the motions of material entities can never be determined with unlimited accuracy. But this does not happen in day to day life. Elaborate.

39. According to Heisenberg's uncertainty principle, the electrons do not have definite position, momentum, energy and time. How our Television pictures are sharp then? Note that in any devices based on cathode ray tube, the pictures are plotted using electrons.
40. Planck's constant is 6.63×10^{-34} JS. Do you know any method of determining it with so much accuracy?
41. Explain how De Broglie's hypothesis would affect our daily life, if Planck's constant were 6.63 J-S instead of 6.63×10^{-34} J-S.
42. Explain how Heisenberg's uncertainty principle would affect our daily life if Planck's constant were 6.63 J.S instead of 6.63×10^{-34} J.S.
43. "*The motion of a subatomic particle can never be determined without disturbing it*". Comment
44. "The uncertainties indicated by Heisenberg's uncertainty principle are not due to erroneous instruments or human errors, but they are inherent in nature and are the law of nature". Comment
45. According to De Broglie's hypothesis, electron is associated with a De Broglie wave. Should its antiparticle also be associated with a De Broglie wave? In what respects the De Broglie waves associated with electron and its antiparticle should differ? The fact is that when electron meets its antiparticle, both annihilate, and a photon is emitted. What must be happening to their De Broglie waves then?
46. Electron diffractometer is equipment similar to Bragg's X ray spectrometer. How does an electron diffractometer work? Where and how it can be used?

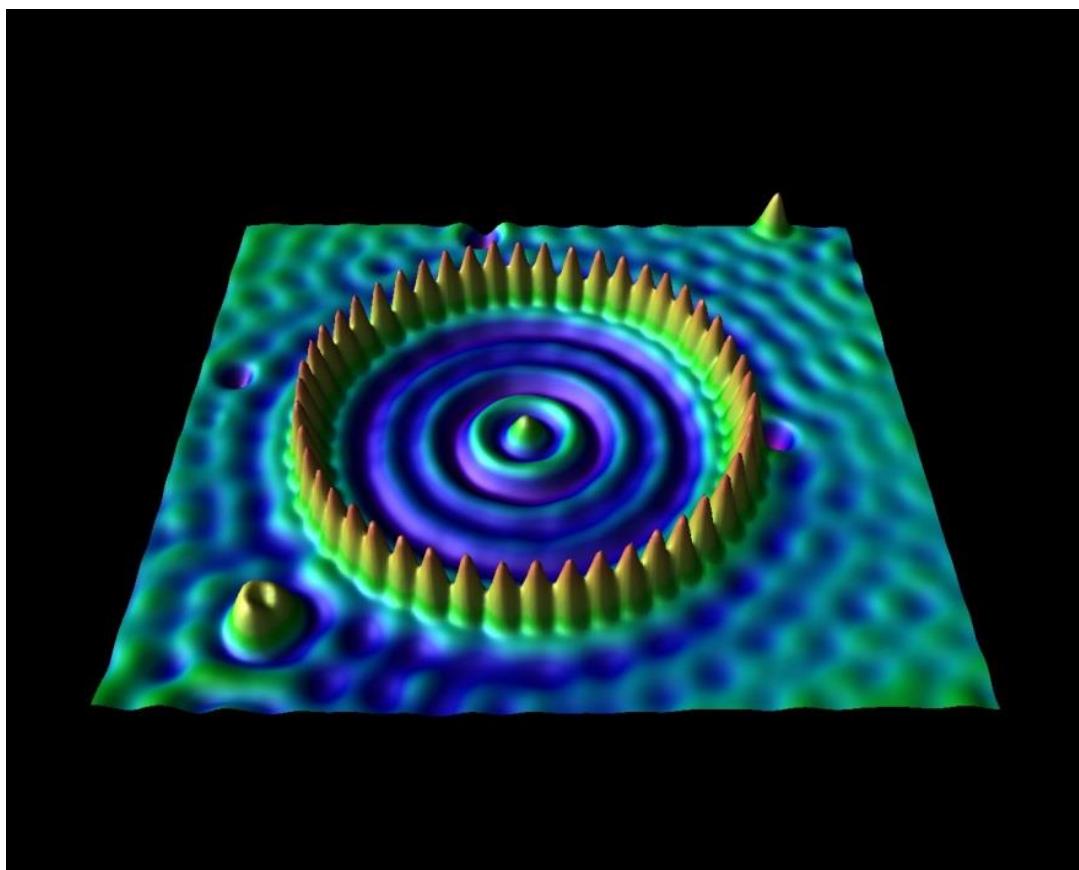
PROBLEMS

A thing is obvious mathematically after you see it

Anonymous

1. Calculate the De Broglie wavelength of proton and deuteron, both having energy 1 keV
2. Using the data of your own choice, prove that De Broglie's hypothesis practically works only for the subatomic particles like electron, but it is not practically applicable to the day to day objects.
3. Calculate the De Broglie wavelength accelerated by the potential of 1 volt, 100 volts and 1000 volts. What happens to De Broglie wavelength of electrons when it is accelerated by increasingly higher potential difference?
4. Use De Broglie's hypothesis to prove that the resolving power of electron microscope is quite higher than that of optical microscope. Does electron behave like particle or wave in electron microscope?
5. Electron cannot exist inside nucleus, because it does so, it's velocity will exceed the speed of light and its minimum energy will be far greater than the binding energy of the nucleus. But, though electron cannot exist inside the nucleus, it can exist inside the atom. Explain this with appropriate calculations. (the size of the atom = 10^{-10} m and the ionization potentials of atoms are typically 3 to 4 eV)

6. We know that electron cannot exist inside the nucleus, because nuclear dimensions (10^{-14} m) are too small to accommodate the electron. At such size of the nucleus its velocity will exceed the speed of light and its minimum energy will be far greater than the binding energy of the nucleus. What should be the minimum size of the nucleus at which electron can exist inside the nucleus? (Given maximum allowed speed in the universe = 3×10^8 m/s and maximum binding energy of the nucleus = 8.8 MeV)
7. As explained above, electron cannot exist inside nucleus. But proton or neutron can exist inside the nucleus. How? (Size of nucleus is 10^{-14} m and mass of proton \cong mass of neutron = 1.67×10^{-27} kg).
8. One reason for non-existence of electron in the nucleus is extremely small dimensions of the nucleus, another reason is the typical value of the Planck's constant = $h = 6.63 \times 10^{-34}$ J.S. At what value of the Planck's constant electron would have existed inside the nucleus, if nuclear dimensions were to remain same i.e. 10^{-14} m?
9. The wavelength of red light is 6330 \AA° . Through what potential difference must an electron be accelerated so that it will acquire this wavelength?



In IBM Almaden Research Center in San Jose, California, the quantum corral was demonstrated in 1993 by Lutz, Eigler, and Crommie using an elliptical ring of 48 iron atoms on a copper surface using the tip of STM. The corral The De Broglie waves inside the nano traps can be easily seen. Matter exhibits wave-like properties at atomic scale

REFERENCE BOOKS

6. Fundamentals of Physics, 9th Edition, Extended, Wiley Resnick, Halliday, Walker,
7. Concepts of Modern Physics, Arthur Beiser, 6th Edition, Tata McGraw Hill
8. Introduction to Quantum Mechanics. - By D. Griffiths Published by Prentice Hall.
9. Quantum Mechanics. - By Ghatak and Loka Nathan Published by Mc. Millan
10. Quantum Mechanics. - By L. I. Schiff.
11. A Text-book of Quantum Mechanics by P.M. Mathews and K. Venkatesan, 2nd Edition, McGraw Hill Education, 2010
12. Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles, 2nd Edition, by Robert Eisberg , Robert Resnick, Wiley
13. Quantum Mechanics: Concepts and Applications, by Nouredine Zettili, 2016, Wiley

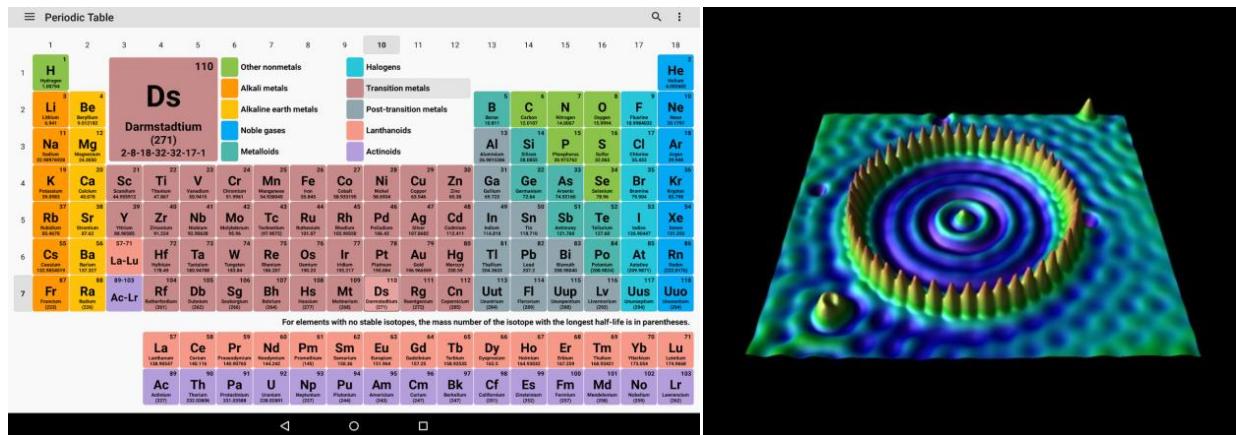
WORLD WIDE WEB

1. <https://quantumphysics.iop.org/>
2. <https://quantumphysicsmadesimple.com/>

CHAPTER 9

Quantum Mechanics and its Applications

(Wave function and Wave equation)



The image on the left shows periodic table of the elements, now understood to a great extent. The electronic configuration of every element in the periodic table is now well understood and found to be consistent with the observed physical and chemical properties of elements. This understanding has led to several applications of various elements and their compounds. Understanding periodic table can be considered to be a greatest triumph of Quantum Mechanics, the theory of matter and energy at atomic scale. The photograph on the right shows a Quantum Corral created by American Physicists in IBM lab. In this Corral, 48 atoms of iron are arranged over the surface of copper. The size of the corral is just 7 nm. This has been done with the help of Scanning tunneling Microscope (STM), an ultra-sophisticated instrument, which itself is based on Quantum laws. The matter waves inside the corral can be conspicuously observed. Thus the atoms can now be ‘seen’ and manipulated. Quantum Mechanics is thus taking the technology towards atomic or a nanometric scale. What is Quantum Mechanics? And what are its foundations?

The answer to this question is in this chapter

Index

9.1 INTRODUCTION

The behavior of atoms, molecules and photons is governed by Quantum laws

9.2 WAVEFUNCTION

Describing De Broglie waves mathematically

9.3 SCHRÖDINGER'S EQUATION

An equation which every acceptable wavefunction must satisfy

9.4 PARTICLE IN A RIGID BOX

Restriction leads to quantization

9.5 PARTICLE IN A NON RIGID BOX

The wavefunctions penetrate the boundaries, even for insufficient energies

9.6 TUNNEL EFFECT AND ITS APPLICATIONS

A particle having energy insufficient to cross a barrier can still tunnel through it

9.7 QUANTUM MECHANICS AND ATOMIC STRUCTURE:

The quantum state of an electron in the atom is governed by four quantum numbers

9.1 INTRODUCTION

The behavior of atoms, molecules and photons is governed by Quantum laws

In chapter 8, we have learnt that, material objects, may it be an electron or may it be a cricket ball, all have wave-like properties. Further, we have also seen that, their wave-like behavior results in to unavoidable uncertainties in determining their motions. We have also seen that, because of extremely small value of Planck's constant, the wave-like behavior of the matter is conspicuous at atomic level only. All this clearly indicates that, the mechanics of subatomic entities needs a different approach, which will consider the wave-like, or in fact, the dual behavior of the matter. This approach, which is based on De Broglie's hypothesis and Heisenberg's uncertainty principle, was developed by many Physicists, in the period 1920-1930. A few of them were Erwin Schrodinger, Werner Heisenberg, Paul Dirac, and Max Born etc, all of them being Nobel laureates. Indeed, two schools of thoughts were developed independently by Erwin Schrodinger (Wave-Mechanics) and Heisenberg-Dirac (Matrix Mechanics). The fabric of Schrodinger's approach has two foundations, one, an equation of motion of subatomic particles, known by the name of its originator; the Schrodinger's equation($H\psi = E\psi$), and second, it's solution, ψ , called wavefunction. As noted in previous chapter, every material object has its own ψ , in which the entire knowledge of motion of the object is rooted. None of the motion in the universe is a motion without constraints. Indeed an absolutely free particle is an idealization. Such restrictions on the motions result in restrictions on ψ and, these further result in to the quantization of physical properties, especially the energy of the particles. In this chapter, we aim to learn this Schrödinger's wave-based approach. At first, we will explore Schrodinger's equation, its solution and it's physical significance. We will then apply this approach to a unique problem in Quantum Mechanics, named 'particle in a box'. This problem essentially represents all motions in the universe, as they are the motions in certain kind of boxes, for ex, an electron the 'atomic box' proton or a neutron in 'nuclear box' or an atom in a 'molecular box' (or even a cricket ball restricted to move within the boundaries of the cricket ground is also an example of particle in the box). The restrictions on such motions lead to quantization...the crux, due to which Quantum Mechanics bears its name. Indeed, an extension of this problem has led to the set of quantum laws behind the electronic configuration of atoms and this further showed a way to the complete understanding of periodic table. The quantum mechanical model of the atom, which is based on four quantum numbers, overcomes almost all limitations of the Bohr's model. Bohr's model accounts only to hydrogen, and not to any many-electron-atom. Bohr's model also fails in explaining certain properties of atomic spectra namely, variation in the intensity of spectral lines and their fine structure. Further, Bohr's model works for individual atom only, and cannot even explain how atoms interact with each other to form the aggregates. Quantum mechanics supersede all these limitations. More importantly, Quantum mechanics has 'gifted' several sophisticated devices to the field of science and technology, a few of which include transistor (Nobel prize in 1956), laser (Nobel prize in 1964), electron microscope (Nobel prize 1986), Scanning Tunneling Microscope (STM, Nobel prize in 1986), SQUID (based on Josephson junctions...Nobel prize 1973) etc. The emerging technologies such as Nanotechnology, Photonics, Spintronics and molecular electronics are all based on Quantum Mechanics...a stream which becomes noticeable at atomic scales. It is worth noting that Quantum Mechanics obeys all the characteristics of an ideal theory, a consistent explanation of the experimentally observed

phenomena, prediction of the new phenomena, and experimental verifications of the underlying principles.

As we shall see further, unlike classical mechanics, quantum mechanics describes the behavior of the matter and energy in the language of probability. For example, in Bohr's model of the atom, we say that the radius of the electron's orbit in the ground state is 0.53 \AA° , but in quantum mechanics, we call it as the most probable radius, indicating that, it may have some other values too. Of course, quantum mechanics does not completely replace classical mechanics, but it appears that classical mechanics is a special case of quantum mechanics.



Max Born (1882 –1970) : He was a German Physicist who made significant contributions in diversified areas of Physics such as optics, solid state physics, relativity and mainly Quantum Mechanics. He was awarded Nobel Prize in Physics in 1954 for his "*fundamental research in Quantum Mechanics, especially in the statistical interpretation of the wave function*". He was also a fellow of prestigious Royal Society. He obtained Ph.D. in University of Gottingen and made his career there itself and there, under his leadership, Göttingen became one of the world's foremost centers for Physics. Some notable Physicists namely Enrico Fermi, Werner Heisenberg (along with whom, he developed matrix representation of quantum mechanics). Friedrich Hund, , Maria Goeppert-Mayer, , Robert Oppenheimer (who was involved in the development of first atomic bomb), Wolfgang Pauli, Paul Dirac , Erwin Schrodinger and Eugene Wigner either worked under him or with him. Max Born also spent a few years of his career in Berlin University and University of Edinburgh. In 1926, Born proposed is probabilistic interpretation of $\psi\psi^*$; ψ being the solution of Schrödinger's equation. A group of all these Physicists developed two equivalent formulations of Quantum Mechanics, namely Matrix Mechanics (Heisenberg) and Wave Mechanics (Schrodinger). Indeed 1900 to 1940 can be considered as golden era of Physics, which witnessed two pillars of Physics, namely Quantum Mechanics and Relativity.

9.2 WAVEFUNCTION

Describing De Broglie waves mathematically

Let us now start evolving the quantum mechanics. We must begin with De Broglie's hypothesis and devise a method to it mathematically. This can be done with the help of $\psi(x, y, z, t)$. As we know, it is ψ whose variations make up a De Broglie wave. In previous chapter we have also noted that, De Broglie waves are probability waves, and a quantity $|\psi|^2$ represents the probability density. We have also seen that ψ itself is neither a probability (as it may be positive as well as negative, and sometimes complex also) and nor any other experimentally measurable quantity. In previous chapter in section (1.10), we have also seen that, once ψ (De Broglie wave) of the particle is known, its entire mechanics (which is based on x, t, p, E) can be described, although it is limited by some unavoidable uncertainties.

In previous chapter, we have seen that, the simplest description of De Broglie waves is based on following expressions, which are equivalent to each other

$$\psi = \psi_o \sin(\omega t - kx) \quad \dots(9.1)$$

and

$$\psi = \psi_o \sin \frac{1}{\hbar} (Et - px) \quad \dots(9.2)$$

As we have seen, the equivalence of both these expressions can be established with the help of some entities such as $k = \frac{2\pi}{\lambda}$, $p = \frac{\hbar}{\lambda}$, $\omega = 2\pi\nu$ and $E = h\nu$

Instead of *sine* function, we may choose *cosine* function, or even a combination of both *sine* and *cosine*, i.e. *exponential* function. Thus another useful expression for describing ψ is

$$\psi = \psi_o e^{-\left(\frac{i}{\hbar}\right)(Et - px)} \quad \dots(9.3)$$

(One can see presence of i in the above expression, which inhibits the realistic and direct interpretation of ψ)

Now, as De Broglie waves are probability waves, and as $|\psi|^2$ represents the probability density, we have

$|\psi|^2$ = Probability of finding the particle per unit length (for 1D De Broglie wave) or per unit volume (for 3D De Broglie wave)

Thus,

$|\psi|^2 dx$ = Probability of finding the particle in a region having length dx

$\int_{-x_1}^{+x_2} |\psi|^2 dx$ = Probability of finding the particle in a region between x_1 and x_2

$\int_{-\infty}^{+\infty} |\psi|^2 dx$ = Total probability of finding the particle in the entire space = 1 $\dots(9.5)$

(This is because, a real particle must exist somewhere in the entire space). Also as $|\psi|^2$ signifies probability density, we expect

$$\int_{-\infty}^{+\infty} |\psi|^2 dx \neq 0 \quad \text{and} \quad \int_{-\infty}^{+\infty} |\psi|^2 dx \neq \infty$$

The expression (9.5) is called as a normalization condition, which every wavefunction must satisfy. The wavefunction which satisfies this condition is called as a normalized wavefunction. An acceptable (well behaved) wavefunction must be normalized or normalizable.

As $|\psi|^2$ represents probability (or as De Broglie waves are probability waves), ψ must satisfy a few additional conditions, so that it can yield meaningful results. These are

- i. ψ must be finite for all values of x . This requirement is because of the fact that ψ is related with the probability, which can never be infinite. For ex, a wave function $\psi = \psi_0 \tan x$ cannot be an acceptable wave function, as it turns out to be infinity for $x = 90$. Similarly $\psi = \psi_0 \sin \frac{1}{x}$ is also not an acceptable wavefunction, as it becomes infinite for $x = 0$
- ii. ψ must be single valued, as we can not specify multiple probabilities of finding the particle at the same point. For ex $\psi = \psi_0 \sin \sqrt{x}$ is not an acceptable wavefunction as, for a given value of x , $\sin \sqrt{x}$ will have two different (\pm) values. Similarly ψ shown in the following graph (Fig 9.1) cannot be an acceptable wavefunction for being multiple valued

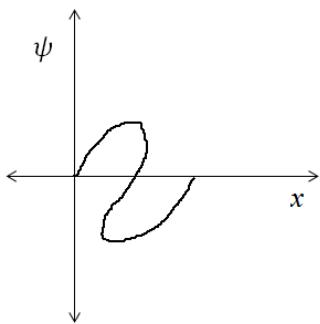


Figure (9.1): Multiple valued ψ

- iii. An acceptable ψ should be continuous, as in real situations, the probability of finding a particle cannot disappear suddenly. In real situations, we must be able to specify the probability of finding the particle at all the points in the region of interest. Consider following two examples of discontinuous wavefunctions

Sr. No.	Formula	Graph	Formula	Graph
I	$\psi(x) = \frac{1}{x-1}$		$\begin{aligned}\psi(x) &= \frac{2}{x^2 - x} \\ &= \frac{2}{x(x-1)}\end{aligned}$	

Table (9.1): Examples of discontinuous ψ

- iv. In addition to ψ , its derivatives $\frac{\partial \psi}{\partial x}, \frac{\partial \psi}{\partial y},$ and $\frac{\partial \psi}{\partial z}$ must also be finite, single valued and

continuous. This condition signifies that De Broglie wave must be ‘smooth’ everywhere, as the probability of finding the particle cannot change abruptly from point to point. The smoothness of the De Broglie wave at a point can be judged by finding the slope there. If the slope (as in Table 9.2) has single value at a point, then the wave can be considered to be smooth. However, if ψ changes abruptly at a given point then there will be two different slopes there (Table 9.2)

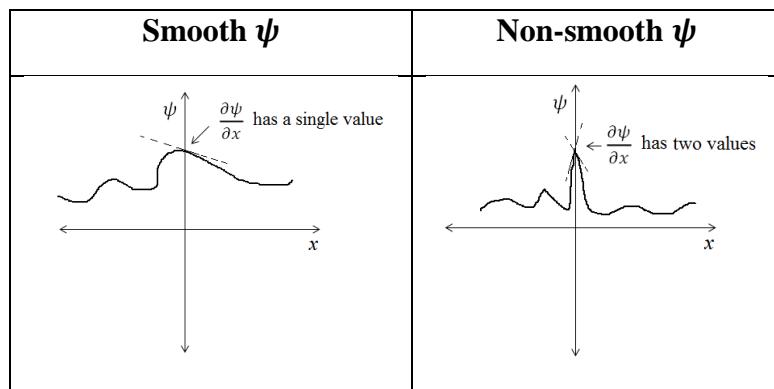
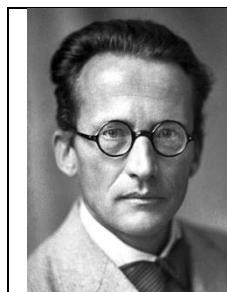


Table (9.2): Smooth and Non-smooth ψ



Erwin Schrodinger (1887-1961): He was an Austrian Physicist, who made significant contributions, not only in Physics (for which he received Nobel prize in 1933), but in such diversified areas as biology, philosophy as well as religion. He served as a Physicist in a few notable universities that included University of Zurich, University of Oxford and University of Vienna.. A few other universities, who offered him a position, included the famous Princeton University and University of Edinburgh. He published several research papers in such diversified areas as Unified Field Theory, concept of gene, an equation known by his name and its applications etc. Indeed his proposition of the concept of gene in his famous book “what is life” resulted in the discovery of structure of DNA in 1953. In early years of his career, he was aware about the quantum concepts developed by contemporary physicists Albert Einstein, Niels Bohr, Arnold Sommerfeld, Wolfgang Pauli etc. Thereby, around 1926, he published his four epoch making papers in which he presented formulation of Schrodinger’s equation, its applications to hydrogen atom, diatomic molecule, harmonic oscillator, rigid rotator, hydrogen atom in electric field, absorption and emission of radiation, and the scattering of radiation by atoms etc. In these papers, he also proved that his approach towards quantum mechanics was very much equivalent to the matrix formulation developed by Heisenberg. It is not an exaggeration to say that, this work itself ‘created’ quantum mechanics, as we know it today. His extensive correspondence with Albert Einstein around 1935 resulted in a famous Schrodinger’s cat thought experiment, which is still a matter of debate. Schrödinger also made serious attempts to formulate grand unified theory, for combining gravity, electromagnetism, weak and nuclear forces. However, this work was left without a success. He was greatly influenced by Vedanta, the Hindu philosophy. Erwin Schrodinger shared with Paul Dirac, the Nobel prize in 1933 for “*for the discovery of new productive forms of atomic theory*”

9.3SCHRÖDINGER'S EQUATION

An equation which every acceptable wavefunction must satisfy

We have seen in details, the nature of matter waves associated with a moving particle, their description based on ψ and their physical interpretation as well. We also know that, in classical physics, we have standard equations that describe a moving particle or a moving (progressive) wave. For ex. we describe a moving particle by a universal equation of motion named Newton's second law, $F = ma$. This equation also describes the motions of mechanical waves, for instance, the sound waves. The propagation of light, which is essentially an electromagnetic wave, requires different kind of equations of motion, named Maxwell's equations. The question now is; can we device a sort of universal equation which will account for the motion of matter waves? Before we answer these questions, let us see, if Newton's law itself can be used to describe the motion of De Broglie waves. This is certainly not possible, as Newton's law assumes that the particle behaves entirely like a particle, that is, it has definite mass and a definite position at an instant. De Broglie's hypothesis and Heisenberg's uncertainty principle collectively indicate that subatomic particles behave as waves and they do not have a definite position at a given instant. Indeed, the subatomic particles have dual character, which means that they behave as wave as well as particles that are like *wavicles*. Thus the Newton's law which considers only particle aspect of the entities cannot account for the motion of the subatomic particles. Even the Maxwell's equations, which completely account for the motions of the electromagnetic waves, cannot describe De Broglie waves, as they consider only the wave aspect of the light. De Broglie waves have dual character and therefore their mathematical description requires a sort of an equation, which will consider both particle as well as the wave aspect of the entities.

Further, as we know, Newton's second laws as well as the Maxwell's equations are the universal equations, in the sense that every moving object in the universe, moving in any circumstances must obey the Newton's law. Similarly all electromagnetic waves moving in any circumstances must obey Maxwell's equations. On similar lines, we need a sort of a universal equation, which will be obeyed by all De Broglie waves (and their associated ψ) under motion. Such equation was developed by an Austrian Physicist named Erwin Schrödinger in 1926. Thus the Schrödinger's equation is a universal equation of motion which must be obeyed by every subatomic particle, the De Broglie wave and the corresponding ψ associated with it. Depending upon the situation such as an electron in various atoms, proton or a neutron in the nucleus, atom in the various molecules, the De Broglie waves and the ψ associated with these entities will be different, however, these entities, their De Broglie waves and corresponding ψ must satisfy a universal equation of motion. This equation is called Schrödinger's equation.

We can arrive at Schrödinger's equation by following number of routes, out of which we will describe two here.

Developing Schrödinger's equation: Approach-I (For exam, follow any one from Approach I and Approach II)

Before we begin, let us discuss what we call as a wave equation

$$\frac{\partial^2 f}{\partial x^2} = \frac{1}{u^2} \frac{\partial^2 f}{\partial t^2} \quad \dots(9.6)$$

The above equation describes all progressive waves such as light, sound, the waves on the water surface etc. If the wave being described by the above equation is light, then the function f represents electric and magnetic field and u corresponds to the speed of light, which is $3 \times 10^8 \text{ m/s}$. If this equation describes sound waves, then f represents the pressure and u represents speed of sound. If it describes waves on the water surface then f is the position of oscillating water molecules on the surface and u represents the speed of such waves. In the next section, we will make a plausible argument, where will argue that, as eqn (9.6) is obeyed by all the progressive waves in the nature, it is to be obeyed by matter waves also. In such case, the function f in the above equation will be ψ - the oscillating quantity in the De Broglie waves and u will be speed of De Broglie waves.

The eqn. (9.6) thus describes variety of waves and therefore can have variety of solutions (including complex ones). These solutions are of the typical form as described below.

$$f = F \left(t \pm \frac{x}{u} \right) \quad \dots(9.7)$$

If above function includes " - " sign, then it describes waves travelling in $+x$ direction and with " + " sign, it describes waves travelling in $-x$ direction. F in the above Eq. can have a form of *sine* or *cosine* or a combination of *sine* and *cosine* i.e. exponential form. If eqn (9.7) is used to describe a wave associated with a free particle (a particle not acted upon by any force, moving on a straight path with constant speed), then wave associated with such particle has constant amplitude (undamped), constant frequency (monochromatic). For such waves, eqn (9.7), takes a form described below.

$$f = f_o e^{-i\omega(t - \frac{x}{u})} \quad \dots(9.8)$$

If the wave is de Broglie wave then, f is ψ , and we have

$$\psi = \psi_o e^{-i\omega(t - \frac{x}{u})}$$

With $\omega = 2\pi\nu$ and $u = \lambda\nu$, we get

$$\psi = \psi_o e^{-i \times 2\pi\nu(t - \frac{x}{\lambda\nu})}$$

$$\psi = \psi_o e^{-i \times 2\pi(vt - \frac{x}{\lambda})}$$

With $E = h\nu$ and $\lambda = \frac{h}{p}$ we get

$$\psi = \psi_o e^{-i \times 2\pi \left(\frac{E}{h} t - \frac{x}{\frac{h}{p}} \right)}$$

Rearranging

$$\psi = \psi_o e^{-i \times \frac{2\pi}{\hbar} (Et - px)} = \psi_o e^{-i \times \frac{1}{\hbar} (Et - px)}$$

Thus

$$\psi = \psi_o e^{-\frac{i}{\hbar} (Et - px)} \quad \dots(9.9)$$

We recall that eqn (9.9) represents a De Broglie wave associated with a free particle moving in $+x$ direction

Differentiating eq (9.10) w.r.t. x we get

$$\frac{\partial \psi}{\partial x} = \left(+ \frac{i}{\hbar} p \right) \psi_o e^{-\frac{i}{\hbar} (Et - px)}$$

And

$$\frac{\partial^2 \psi}{\partial x^2} = \left(+ \frac{i}{\hbar} p \right) \times \left(+ \frac{i}{\hbar} p \right) \psi_o e^{-\frac{i}{\hbar} (Et - px)}$$

Thus

$$\frac{\partial^2 \psi}{\partial x^2} = \left(- \frac{p^2}{\hbar^2} \right) \psi \quad \dots(9.10)$$

Thus

$$\begin{aligned} \frac{\partial^2 \psi}{\partial x^2} &= \left(- \frac{p^2}{\hbar^2} \right) \psi \\ \Rightarrow p^2 \psi &= -\hbar^2 \frac{\partial^2 \psi}{\partial x^2} \end{aligned} \quad \dots(9.11)$$

Differentiating eqn (9.9) w.r.t. t , we get

$$\begin{aligned} \frac{\partial \psi}{\partial t} &= \left(- \frac{i}{\hbar} E \right) \psi_o e^{-\frac{i}{\hbar} (Et - px)} = \left(- \frac{i}{\hbar} E \right) \psi \\ \Rightarrow i\hbar \frac{\partial \psi}{\partial t} &= E\psi \end{aligned} \quad \dots(9.12)$$

We know that, for particle in motion, the total energy E , which is the sum of kinetic and potential energy, is conserved. Thus we have

$$\begin{aligned} E &= KE + PE \\ E &= \frac{1}{2} mv^2 + V(x, t) \end{aligned}$$

$$\Rightarrow E = \frac{p^2}{2m} + V(x, t) \quad \dots(9.13)$$

Note that eqn (9.13) represents law of conservation of energy. Multiplying both sides by ψ , we get

$$E\psi = \frac{p^2}{2m}\psi + V(x, t)\psi$$

Substituting p^2 and E from Eqns (9.11) and (9.12) respectively, we get

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V(x, t)\psi \quad \dots(9.14)$$

In three dimensions,

$$\begin{aligned} i\hbar \frac{\partial \psi}{\partial t} &= -\frac{\hbar^2}{2m} \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} \right) + V(x, t)\psi \\ \Rightarrow i\hbar \frac{\partial \psi}{\partial t} &= -\frac{\hbar^2}{2m} \left(\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} \right) + V(x, t)\psi \\ i\hbar \frac{\partial \psi}{\partial t} &= -\frac{\hbar^2}{2m} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) \psi + V(x, t)\psi \\ \Rightarrow i\hbar \frac{\partial \psi}{\partial t} &= -\frac{\hbar^2}{2m} \nabla^2 \psi + V(x, t)\psi \end{aligned} \quad \dots(9.15)$$

Where ∇^2 represents a Laplacian operator and is given by

$$\nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right)$$

Eqn (9.15) is called as Schrödinger's equation in time dependent form. Note that for developing this Schrödinger's eqn, we have started with Eqn (9.9), which essentially represents an undamped and monochromatic De Broglie wave associated with a free particle. Note that for a free particle, the potential energy is either constant or zero. Thus the expressions (9.11) and (9.12) for $p^2\psi$ and $E\psi$ are applicable to free particle only. On the contrary, the statement of the law of conservation of energy expressed in eqn (9.13) accounts for both types of particles; a free particle ($V = \text{const}$) and a particle moving under some constraints {a bounded particle, with $V = V(x, t)$ }. How can we substitute $p^2\psi$ and $E\psi$ in eqns (9.11) and (9.12) (applicable to a free particle) in the eqn (9.13) applicable to a free as well as bounded particle? As such, this has no justification. Thus, to test whether the Schrödinger's eqn expressed in eqn (9.15) is really valid to a bounded particle moving under some constraints {with $V =$

$V(x, t)\}$ or not, we can follow only an empirical method. We can apply this Schrödinger's eqn to variety of bounded particles moving under different constraints with different forms of $\{V = V(x, t)\}$, obtain its solutions and test them experimentally. It can be noted that, this has been done several times. Physicists have applied Schrödinger's eqn to variety of bounded particles and it has been noticed that its solutions match with experimental observations in all such situations. All this leads to a conclusion that Schrodinger's eqn is a postulate/principle which has been found valid till now. Also note that for developing Schrodinger's eqn., we have used law of conservation of energy. Thus Schrödinger's eqn is embodied in the form of law of conservation of energy. The law of conservation of energy is one of the most fundamental principle of Physics which cannot be derived from anything else. Thus Schrödinger's eqn is one of the fundamental principles of Physics which cannot be derived from anything else. The mathematical exercise that we have carried out here is not a derivation, but one of the ways to formulate the Schrödinger's eqn.

Schrödinger's eqn, time independent (steady state) form:

In the Schrödinger's eqn expressed in (9.15), the term $V(x, t)$ represents the potential energy. Potential energy signifies the 'boundedness' of the particle. In fact a particle, whose motion is restricted, is always acted upon by some force. Force and potential energy are related by following expressions

$$V = \int F dx \text{ or } F = \frac{dV}{dx}$$

We have also seen that, potential energy in eqn (9.15) is a function of x as well as t . However in some situations the potential energy may be a function of only position and it may be independent of time. For ex. the potential energy of an electron in an atom is given by

$$V = -\frac{1}{4\pi\epsilon_0} \frac{Ze^2}{r} \quad \dots(9.16)$$

This potential energy is a function of only r and not t . Further, the potential energy of a harmonic oscillator is given by

$$V = \frac{1}{2} kx^2 \quad \dots(9.17)$$

Here also potential energy is independent of time. In such situations, the Schrödinger's eqn can be reduced to a simpler version which is called as Schrödinger's time independent (steady state) equation. This is shown below.

The solution of Schrödinger's eqn is given in eqn (9).

$$\psi = \psi_o e^{-\frac{i}{\hbar}(Et - px)}$$

$$\Rightarrow \psi = \psi_o e^{(-\frac{iE}{\hbar})t} \times e^{(\frac{ip}{\hbar})x}$$

The above eqn contains space dependent as well as time dependent parts separated from each other

$$\Rightarrow \psi = \psi' e^{\left(-\frac{iE}{\hbar}\right)t} \quad \dots(9.17)$$

Where, $\psi' = \psi_o e^{\left(\frac{ip}{\hbar}\right)x}$ is the space dependent part. Now let us recall Schrödinger's time dependent eqn

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V(x, t)\psi$$

Substituting ψ from eqn (9.17) in above eqn we get

$$\begin{aligned} i\hbar \frac{\partial}{\partial t} \left(\psi' e^{\left(-\frac{iE}{\hbar}\right)t} \right) &= -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \left(\psi' e^{\left(-\frac{iE}{\hbar}\right)t} \right) + V(x, t) \left(\psi' e^{\left(-\frac{iE}{\hbar}\right)t} \right) \\ \Rightarrow i\hbar \psi' \frac{\partial}{\partial t} \left(e^{\left(-\frac{iE}{\hbar}\right)t} \right) &= -\frac{\hbar^2}{2m} \left(e^{\left(-\frac{iE}{\hbar}\right)t} \right) \frac{\partial^2 \psi'}{\partial x^2} + V(x, t) \left(\psi' e^{\left(-\frac{iE}{\hbar}\right)t} \right) \\ \Rightarrow i\hbar \psi' \frac{\partial}{\partial t} \left(e^{\left(-\frac{iE}{\hbar}\right)t} \right) &= -\frac{\hbar^2}{2m} \left(e^{\left(-\frac{iE}{\hbar}\right)t} \right) \frac{\partial^2 \psi'}{\partial x^2} + V(x, t) \left(\psi' e^{\left(-\frac{iE}{\hbar}\right)t} \right) \\ \Rightarrow i\hbar \psi' \left(-\frac{iE}{\hbar} \right) \left(e^{\left(-\frac{iE}{\hbar}\right)t} \right) &= -\frac{\hbar^2}{2m} \left(e^{\left(-\frac{iE}{\hbar}\right)t} \right) \frac{\partial^2 \psi'}{\partial x^2} + V(x, t) \left(\psi' e^{\left(-\frac{iE}{\hbar}\right)t} \right) \end{aligned}$$

Cancelling the common term from both the sides

$$E\psi' = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi'}{\partial x^2} + V(x, t)\psi'$$

For convenience, we choose a notation ψ for ψ' . Thus we have

$$E\psi = -\frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} + V(x, t)\psi \quad \dots(9.18)$$

Above eqn contains only space dependent terms. The time dependent terms have been eliminated. The eqn is thus called as Schrödinger's time independent (steady state) equation.

Developing Schrödinger's equation: Approach II

We have seen that Schrödinger's equation can be developed but cannot be derived. The approach that we discuss now is also not a derivation, but one of the ways to arrive at Schrödinger's equation.

Let us recall the equation of a progressive wave (Eqn (9.6)) .

$$\frac{\partial^2 f}{\partial x^2} = \frac{1}{u^2} \frac{\partial^2 f}{\partial t^2}$$

In three dimensions

$$\nabla^2 f = \frac{1}{u^2} \frac{\partial^2 f}{\partial t^2} \quad \dots(9.19)$$

As discussed earlier ∇^2 is a Laplacian operator which is given by

$$\nabla^2 = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right)$$

Eqn (9.19) is obeyed by all waves in the nature such as light, sound, waves on string or waves on water surface etc. Let us make a plausible argument now. We say that, as Eq (9.19) is applicable to all the waves in the nature, even De Broglie waves may also obey it. For applying Eqn (9.19) to De Broglie waves, we replace f by ψ . Further, the term u now represents phase velocity of De Broglie waves. Thus

$$\nabla^2 \psi = \frac{1}{u^2} \frac{\partial^2 \psi}{\partial t^2} \quad \dots(9.20)$$

The above equation can very well be treated as Schrödinger's equation. However, it is in raw form, in the sense that in it, both ψ as well as u represent wave properties. We need to bring this eqn in a form which will include wave as well as particle properties of De Broglie waves. In another language, somehow, we need to incorporate following property of De Broglie waves in Eqn (9.20)

$$\lambda = \frac{h}{p}$$

For this purpose, we use following approach. The most basic eqn, which represents a simplest wave is

$$\psi = \psi_0 \sin(\omega t - kx) \quad \dots(9.21)$$

Eqn (9.21) can even be expressed in cosine form. As De Broglie waves are wavegroups, they can be best represented by an exponential function, which is the superposition of *sine* and *cos* functions. Thus

$$\psi = \psi_0 e^{-i(\omega t - kx)} \quad \dots(9.22)$$

We now separate the space dependent and time dependent parts of above eqn

$$\psi = \psi_0 e^{+ikx} e^{-i\omega t}$$

We can denote the space dependent term $(\psi_0 e^{+ikx})$ by ψ'_0 . Thus, we have

$$\psi = \psi'_0 e^{-i\omega t} \quad \dots(9.23)$$

Differentiating ψ w.r.t time, we get

$$\frac{\partial \psi}{\partial t} = (-i\omega) \psi'_0 e^{-i\omega t}$$

Differentiating once again

$$\frac{\partial^2 \psi}{\partial t^2} = (-i\omega)(-i\omega) \psi'_0 e^{-i\omega t} = -\omega^2 \psi \quad \dots(9.24)$$

Eqn (9.24) helps us to eliminate the time dependent term in eqn (9.20). Substituting

$$\nabla^2 \psi = -\frac{\omega^2}{u^2} \psi$$

Using the identities, $\omega = 2\pi\nu$ and $u = \lambda\nu$, we get

$$\begin{aligned} \nabla^2 \psi &= -\frac{(2\pi\nu)^2}{(\lambda\nu)^2} \psi \\ \Rightarrow \nabla^2 \psi &= -\left(\frac{2\pi}{\lambda}\right)^2 \psi \end{aligned} \quad \dots(9.25)$$

Now Eqn (9.25) has acquired a form, where we can easily substitute De Beoglie's eqn $(\lambda = \frac{h}{p})$.

Thus we have

$$\begin{aligned} \nabla^2 \psi &= -\left(\frac{2\pi}{\frac{h}{p}}\right)^2 \psi \\ \Rightarrow \nabla^2 \psi &= -\left(\frac{p}{\frac{h}{2\pi}}\right)^2 \psi \\ \Rightarrow \nabla^2 \psi &= -\frac{p^2}{\hbar^2} \psi \end{aligned} \quad \dots(9.26)$$

Note that the eqns (9.20, 9.25 and 9.26); all of them represent Schrödinger's equations in various forms such as u , ω , λ and p . Of course the most convenient form is eq (9.26), in terms of p . We now aim to express the eqn (9.27) in terms of physical quantity, called energy (E) which is more

fundamental than momentum (p). For a particle in motion, the total energy (E), which is the sum of kinetic and potential energies, is always conserved. We have

$$E = \frac{1}{2}mv^2 + V(x, t)$$

$$\Rightarrow E = \frac{p^2}{2m} + V(x, t)$$

$$\Rightarrow p^2 = 2m\{E - V(x, t)\}$$

Thus now p^2 in Eqn () can be conveniently expressed in terms of energy E . Substituting

$$\nabla^2\psi = -\frac{2m\{E - V(x, t)\}}{\hbar^2}\psi$$

Rearranging,

$$-\frac{\hbar^2}{2m}\nabla^2\psi + V(x, t)\psi = E\psi \quad \dots(9.27)$$

Eqn (9.28) is called as Schrödinger's equation in time independent (steady state) form. Note that, if the above equation is compared with the statement of law of conservation of energy, then the terms $E\psi$, $V(x, t)\psi$ and $\frac{\hbar^2}{2m}\nabla^2\psi$, correlate to total energy (E), potential energy{ $V(x, t)$ } and the kinetic energy. We thus conclude that Schrödinger's equation is basically an equation embodied in the form of law of conservation of energy, a fundamental principle in Physics. Thus we note that Schrodinger's equation cannot be derived from any other principle. Schrödinger's equation itself is a fundamental principle (which cannot be derived from anything else).

We can easily formulate Schrödinger's time dependent equation eqn now. Recall Eqn (9.23), which is

$$\psi = \psi'_o e^{-i\omega t}$$

Differentiating w.r.t. time, we get

$$\frac{\partial\psi}{\partial t} = (-i\omega)\psi'_o e^{-i\omega t}$$

$$\Rightarrow \frac{\partial\psi}{\partial t} = (-i\omega)\psi$$

Using identities $\omega = 2\pi\nu$ and $E = h\nu$, we get

$$\frac{\partial\psi}{\partial t} = \left(-i2\pi\frac{E}{\hbar}\right)\psi$$

$$\Rightarrow \frac{\partial \psi}{\partial t} = \frac{1}{i\hbar} E\psi$$

$$\Rightarrow i\hbar \frac{\partial \psi}{\partial t} = E\psi \quad \dots(9.28)$$

Substituting in eqn (9.27), we get

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V(x, t)\psi = i\hbar \frac{\partial \psi}{\partial t} \quad \dots(9.29)$$

Eqn. (9.29) is called as Schrödinger's time dependent equation. We can rewrite eqn (9.29) in the following form

$$\left\{ -\frac{\hbar^2}{2m} \nabla^2 + V(x, t) \right\} \psi = E\psi$$

This helps us to express the Schrödinger's equation in a still compact form, given by

$$H\psi = E\psi \quad (\text{time independent form}) \quad \dots(9.30)$$

And

$$H\psi = i\hbar \frac{\partial \psi}{\partial t} \quad (\text{time dependent form}) \quad \dots(9.31)$$

Where H represents Hamiltonian operator and is given by

$$H = \left\{ -\frac{\hbar^2}{2m} \nabla^2 + V(x, t) \right\} \quad \dots(9.32)$$

Schrödinger's equation in time independent (steady state) and time dependent form: a closer look

Let us once again recall both the equations

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V(x, t)\psi = E\psi \quad (\text{time independent form}) \quad \dots(9.30)$$

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V(x, t)\psi = i\hbar \frac{\partial \psi}{\partial t} \quad (\text{time dependent form}) \quad \dots(9.31)$$

Consider the potential energy of an electron moving around the nucleus. We have

$$V(r) = -\frac{1}{4\pi\epsilon_0} \frac{Ze^2}{r} \quad \dots(9.33)$$

Further, the potential energy of a harmonic oscillator is given by

$$V(x) = \frac{1}{2}kx^2 \quad \dots(9.34)$$

As the potential energies in eqn (9.33) and (9.34) are only space dependent and time independent, we can solve such problems very well with Schrödinger's time independent equation. As this equation does not involve any time derivative of ψ , while solving it, we do not perform integration over time. As the integration is carried over only space coordinates, the solution $\psi(x, y, z)$ is a function of only space coordinates. Thus the De Broglie waves and the corresponding quantum states are 'stationary' w.r.t. time.

Now let us consider the motion of an electron in an atom placed in a time dependent electric or magnetic field. In such cases, the potential energy {given in eqn. (9.33)} will be a function of time also. Further, consider the motion of a harmonic oscillator, where the spring is held in a furnace. The elastic constant of the spring and consequently the potential energy will then change with time. Such motions are attempted with Schrödinger's time dependent equation. As the equation contains both space as well as time derivatives, while solving, the integrations are performed over both space and time coordinates. The solution $\psi(x, y, z, t)$ is thus the function of both space and time coordinates. The quantum states described by such wavefunctions, the corresponding De Broglie waves and the energies are no longer stationary. They evolve with time. Also note that, if we take a 'snapshot' of such states at a given instant, they are frozen and then such snapshots represent the steady state De Broglie waves, which are the solutions of the Schrödinger's time independent equation.

9.4 PARTICLE IN A RIGID BOX (INFINITE POTENTIAL WELL)

Restriction leads to quantization

In chapter (8) and section (8.14), we have attempted the problem of motion of a particle in rigid box using first principles of quantum mechanics. There we have learnt that, the energy of a trapped particle is quantized and the ground state energy of such particles cannot be brought down to absolute zero. Now we aim to approach the same problem by using Schrödinger's equation. Indeed, solving such equation for variety of problems in subatomic regime itself is what is called as 'fundamental' method of quantum mechanics. We have also noticed there that, all motions in the nature are in a way, the motions in 'boxes' (such as electron in atom, nucleon in nucleus or an atom in molecule or even a cricket ball on the ground). However, such motions are quite intricate and solving Schrödinger's equation for such motions requires an elaborate and laborious mathematics. In section (8.14) we have also seen that, we have simplified the problem by making some assumptions such as infinitely rigid walls, one dimensional motion and zero potential energy. We continue here with the same assumptions. As we shall notice here, even such an oversimplified problem involves a considerable mathematics.

This problem has been already defined in section (8.14). Refer fig (8.12), which is redrawn here (Fig 9.2). We recall that we have to analyze a motion of a particle in an infinitely rigid box (infinitely deep potential well). We also recall that the potential energy (V) of the particle at $x = 0$ and $x = \infty$ is infinite, while inside the box (i. e. $0 < x < L$) it is zero.

As the potential well, as defined in above way does not change with time, we will attempt

this problem by using Schrödinger's time independent (steady state) equation.. We have

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V(x, t)\psi = E\psi$$

As the motion is one dimensional, ∇^2 reduces to $\frac{\partial^2}{\partial x^2}$ and due to the same reason (one

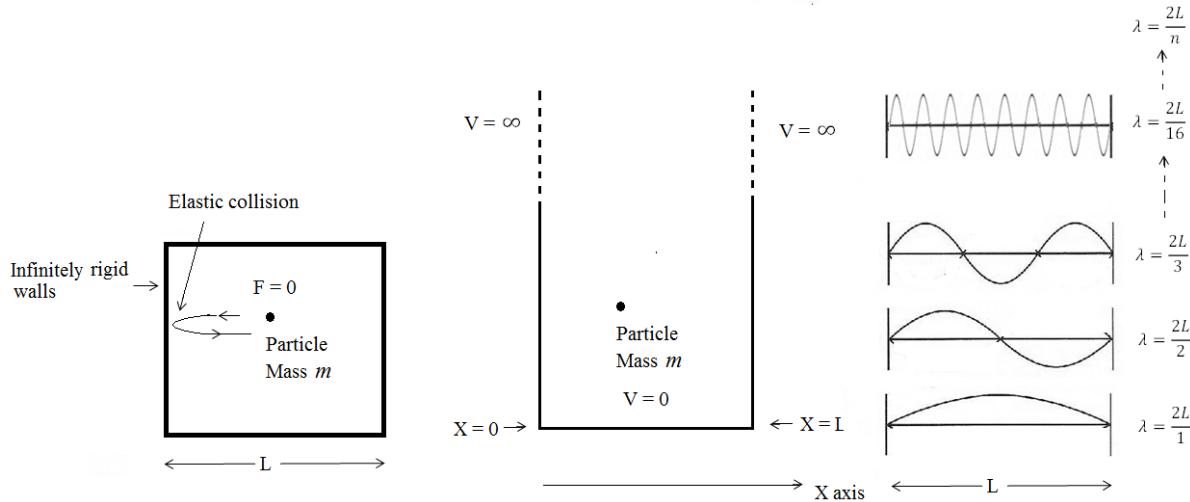


Figure (9.2): Particle in rigid box; the box, the potential well and wavefunctions

dimensionality) $\frac{\partial^2}{\partial x^2}$ becomes to $\frac{d^2}{dx^2}$. Further, potential energy inside the box is zero. Thus, we have

$$-\frac{\hbar^2}{2m} \frac{d^2\psi}{dx^2} + 0 \times \psi = E\psi$$

Rearranging,

$$\frac{d^2\psi}{dx^2} + \left(\frac{2mE}{\hbar^2}\right)\psi = 0 \quad \dots(9.34)$$

We recall that the walls of the box being infinitely rigid, the collisions of the particle with the walls are elastic and thus the energy E is constant. The other terms in the bracket in the above equation are constant. Thus we choose

$$\left(\frac{2mE}{\hbar^2}\right) = \text{constant} = k^2 \quad \dots(9.35)$$

Thus

$$\frac{d^2\psi}{dx^2} + k^2\psi = 0 \quad \dots(9.36)$$

(Note that we represent the bracket by k^2 and not k , as, with k^2 , the differential eqn (9.36), acquires a standard form, whose solution can then be easily found)

Eqn (9.36) is one of the standard equations in mathematics and it has standard solutions, which are,

$$\psi = \psi_0 \sin(kx + B) \quad \dots(9.37)$$

OR

$$\psi = A \sin kx + B \cos kx \quad \dots(9.38)$$

OR

$$\psi = A e^{ikx} + B e^{-ikx} \quad \dots(9.39)$$

All above solutions can be verified by substituting them in eqn (9.36). We can proceed ahead with any of these solutions, but we choose (9.37) which is simpler

$$\psi = \psi_0 \sin(kx + B) \quad \dots(9.37)$$

Above solution ψ appears to be a well behaved (acceptable) wavefunction as it is finite, continuous and single valued. However, for ψ to describe the solution of this problem aptly, we need to test it for the ‘boundary conditions of the problem’. As the walls of the box are infinitely rigid, we expect that the probability (ψ^2) of finding the particle at $x = 0$ and $x = L$ be zero. Thus we expect $\psi = 0$ at $x = 0$ and $x = L$ (these are called as boundary conditions). Thus the constants ψ_0 , k and B in eqn (9.37) cannot take arbitrary values. Their values should be evaluated by applying the boundary conditions.

The first boundary condition is $\psi = 0$ at $x = 0$. Applying this to ψ in Eqn (9.37)

$$0 = \psi_0 \sin(k \times 0 + B)$$

$$\Rightarrow 0 = \psi_0 \sin B \Rightarrow \text{either } \psi_0 = 0 \text{ or } \sin B = 0$$

However, ψ_0 cannot be taken zero as then the wavefunction disappears at all values of x . Now $\sin B = 0$ requires either $B = 0, \pm\pi, \pm 2\pi, \pm 3\pi, \dots$. We proceed with $B = 0$ (being the simplest one). Thus

$$\psi = \psi_0 \sin kx \quad \dots(9.38)$$

Second boundary condition is $\psi = 0$ at $x = L$. Applying to Eqn (9.38), we get

 Quantization!	$0 = \psi_0 \sin kL \Rightarrow \text{either } \psi_0 = 0 \text{ or } \sin kL = 0$ Once again ψ_0 cannot be taken zero due to reason explained above, we thus have $\sin kL = 0 \Rightarrow kL = n\pi \Rightarrow k = \frac{n\pi}{L}$ $\dots(9.39)$
--	--

Here n is an integer with values $0, \pm 1, \pm 2, \pm 3 \dots$. Principally we may start with $n = 0$ but the problem does not permit, as per Eqn. (9.39) with $n = 0$, we have $k = 0$. Eqn (9.38) indicates that this will give $\psi = 0$ for any value of x indicating absence of particle. Thus we have $n = \pm 1, \pm 2, \pm 3 \dots$ etc. As the values of n are restricted to only integers, it is called '**quantum number**'.

Substituting $k = \frac{n\pi}{L}$ in eqn (9.38), we get

$$\psi = \psi_0 \sin \frac{n\pi}{L} x \quad \dots(9.40)$$

It can be seen that ψ in Eqn (9.40) satisfies both the boundary conditions. However, this is not enough, as an acceptable ψ has to satisfy normalization condition also (refer eqn. (9.5)). Thus

$$\int_{-\infty}^{+\infty} \psi^2 dx = \int_0^L \psi^2 dx = 1 \quad \dots(9.41)$$

Substituting ψ ,

$$\begin{aligned} & \int_0^L \left(\psi_0 \sin \frac{n\pi}{L} x \right)^2 dx = 1 \\ & \Rightarrow \psi_0^2 \int_0^L \left(\sin \frac{n\pi}{L} x \right)^2 dx = 1 \end{aligned}$$

It can be shown that

$$\int_0^L \left(\sin \frac{n\pi}{L} x \right)^2 dx = \frac{L}{2}$$

Substituting back,

$$\psi_0^2 \times \frac{L}{2} = 1$$

Giving

$$\psi_0 = \sqrt{\frac{2}{L}}$$

Substituting this ψ_0 in Eqn (9.40), we get

$$\psi_n = \sqrt{\frac{2}{L}} \sin \frac{n\pi}{L} x \quad \dots(9.41)$$

Eqn (9.41), gives the complete wavefunction (De Broglie wave) of the particle trapped in the rigid box. The wavefunction in the eqn (9.37) is called as the general solution because all the

constants ψ_0 , k and B are unknown. Eqn (9.41) represents the particular solution, where all the constants ψ_0 , k and B are evaluated by considering the characteristics of the problem.

Now recall eqn (9.39) We have seen that $k \left(= \frac{n\pi}{L} \right)$ is quantized due to the boundary condition $\psi = 0$ at $x = L$. We have already defined k^2 is eqn (9.35). Equating k^2 in eqn (9.35) and (9.39), we get

$$k^2 = \left(\frac{2mE}{\hbar^2} \right) = \left(\frac{n\pi}{L} \right)^2$$

Rearranging,

$$E_n = \frac{n^2 \hbar^2}{8mL^2} \quad \dots(9.42)$$

Recall that the nature of boundary condition, $\psi = 0$ at $x = L$ requires n to be a quantum number, which is allowed to take only integer values excluding zero. Thus we see that the energy of a particle in a rigid box is quantized. By assigning various values to n we can compute a few energy levels of the particle. Refer Fig (9.3). We note that the energy levels are discrete (quantized) and unevenly spaced(due to n^2).

We now approach towards quantization of the momentum and the wavelength. We have

$$\text{Total energy } E = KE + PE$$

$$\Rightarrow E = \frac{1}{2}mv^2 + V(x, t)$$

As in our problem, $V(x, t) = 0$,

$$\Rightarrow E = \frac{p^2}{2m}$$

$$\Rightarrow p = \sqrt{2mE}$$

Substituting E from eqn (9.42),

$$p = \sqrt{2m \left(\frac{n^2 \hbar^2}{8mL^2} \right)} \\ \Rightarrow p_n = \pm \frac{n\hbar}{2L} \quad \dots(9.43)$$

Thus along with energy, momentum is also quantized. The \pm sign indicates that momentum has direction. We also have

$$\lambda = \frac{\hbar}{p}$$

Substituting p from eqn (9.43),

$$\lambda = \frac{h}{\left(\frac{nh}{2L}\right)} = \frac{2L}{n} \quad \dots(9.44)$$

Thus wavelength is also quantized. Recall that the formulae for energy, momentum and the wavelength that we derived here are the same as we have obtained in section (8.14) in chapter (8), where we attempted this problem by using the first principles of quantum mechanics. This confirms that an approach based on Schrodinger's eqn is consistent with the first principles in quantum mechanics, such as De Broglie's hypothesis and the Heisenberg's uncertainty principle. This also a confirmatory test of all these principles, as all of them is consistent with each other. Also recall that the formula for the wavelength in eqn (9.44) is similar to the wavelengths of standing waves produced on a string of length L tied between two rigid supports. In section (8.3) and (8.14), we have also seen how such standing waves appear pictorially.

A closer look at the second boundary condition ($\psi = 0$ at $x = L$) and the corresponding quantization of k, E, p and λ clearly indicates that, a sort of restriction on the motion always leads to quantization. Note that the boundary condition ($\psi = 0$ at $x = L$ or $x \geq L$) signifies a restriction on the motion. For a free particle, there is no boundary condition and therefore there is no quantization. Thus the formal approach of quantum mechanics based on Schrödinger's equation also confirms that

The energy of free particle is never quantized. But the energy of a particle bounded (in any way) is always quantized

Till now, we have described the energy and momentum and the wavelength of a particle in a rigid box, however, the nature of De Broglie waves and the probable positions in various quantum states are yet to be described. Recall the eqn (9.41) for ψ_n . We can evaluate $\psi_n, E_n, p_n, \lambda_n$ and the corresponding quantum states by choosing $n = 1, 2, 3, \dots, etc$

Refer Fig (9.4). We can obtain the graphs of ψ and ψ^2 in various quantum states by assigning various values to x

[such as $(0, \frac{L}{2}, L)$ for $n = 1$, $(0, \frac{2L}{4}, \frac{3L}{4}, \frac{4L}{4})$ for $n = 2$, $(0, \frac{L}{6}, \frac{2L}{6}, \frac{3L}{6}, \frac{4L}{6}, \frac{5L}{6}, \frac{6L}{6})$ for $n = 3$ etc.]. The Figure (9.4) shows the graphs for $n = 1, 2, 3, \text{ and } 12$. Though ψ_1, ψ_2, ψ_3 and ψ_{12} give us the De Broglie waves, as we have seen earlier, it is $\psi_1^2, \psi_2^2, \psi_3^2$ and ψ_{12}^2 which have significance.

$$E_4 = \frac{16h^2}{8mL^2}$$

$$E_3 = \frac{9h^2}{8mL^2}$$

$$E_2 = \frac{4h^2}{8mL^2}$$

$$E_1 = \frac{h^2}{8mL^2}$$

Fig (9.3): Discrete energy levels of a particle in rigid box

We also recall that these ‘squared’ ψ represent the probabilities of finding the particle at various positions in the potential well. We observed that in the lowest energy state, i.e. the ground state, the particle is most likely to be found at the center of the box, and the probability of finding the particle there is $\frac{2}{L}$. This result contradicts with classical physics. We have noted that the energy of the particle in a given quantum state is constant. Thus the speed in a given quantum state is constant. For a particle moving inside the box with uniform speed, all positions are equally probable. Thus we see that in ground state, quantum mechanics disagrees with classical physics (day to day physics). In the next state (corresponding to $n = 2$), we find that the particle is most likely to be found at $\frac{L}{4}$ and $\frac{3L}{4}$, however at $\frac{L}{2}$, which is the position of maximum probability in the ground state, there is no chance of finding the particle in next quantum state.

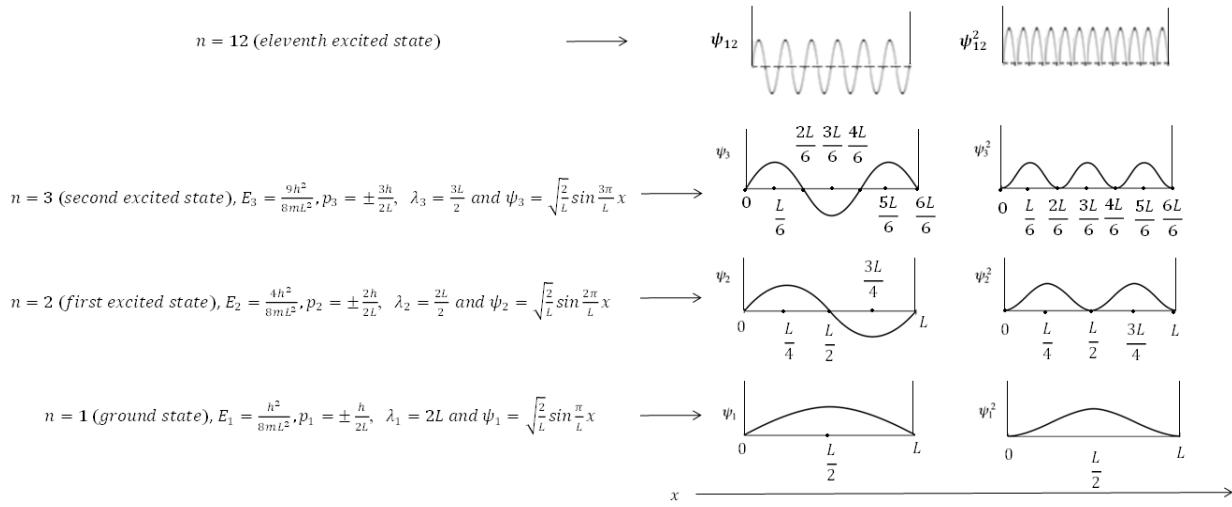


Fig (9.4) Stationary states of the particle in rigid box

Once again we get a result which cannot be perceived using common sense. Of course we have seen that quantum mechanics disagrees with the common sense in many situations (such as ‘particles’ have ‘wavelike’ properties, their motion as are not completely predictable, their energies are quantized so on...).

A careful look at the graphs of ψ_n^2 Vs x for various values of n clearly indicates that the number of peaks in the graph is exactly equal to the quantum number of state. For ex. for $n = 1, 2$ and 3 the number of peaks in the graphs of ψ_1^2 , ψ_2^2 and ψ_3^2 Vs x are 1 , 2 and 3 respectively. A logical extension of the situation suggests that for $n = 12$, there will be 12 peaks, for $n = 120$ there will be 120 peaks, for $n = 10^{12}$, there will be 10^{12} peaks a so on! We thus note that for an extremely large value of the quantum number, the number of peaks will be so large that, practically, all positions will become equally probable. As discussed earlier, classical mechanics expects all positions to be equally probable, if the energy (and the speed) is uniform. We thus conclude that quantum mechanics approaches classical mechanics when the quantum

states correspond to extremely large quantum number. As we shall see further, the subatomic particles are characterized by low quantum numbers, but the quantum numbers of objects in day to day life are extremely large. Further, we also note that the quantization is conspicuous only for low quantum numbers. For ex there is a clear difference between the quantum states corresponding to say $n = 2$ and 3 or 5 and 6 etc but for extremely large quantum numbers, for ex. the difference between say $n = 10^{40}$ and $n = (10^{40} + 1)$ is negligible. Thus quantization becomes unnoticeable for states of extremely large quantum numbers. All this discussion leads to three conclusions

- i. Quantum mechanics and classical mechanics give same result for the day to day objects. This means that for the analysis of day to day objects, the mathematically intricate quantum mechanics can be avoided, and a simpler approach based on classical physics can be followed.
- ii. Physics thus differentiates the world around us in two categories, the world of subatomic particles, which is entirely governed by quantum mechanics and the day to day world governed by classical physics
- iii. Does this really mean that the Physics of atomic world and day to day world are exclusively different? Someone, who knows adequate physics, will be certainly uncomfortable with this situation. But, he need not be; because, we have proven that quantum mechanics gives the same results as that of classical physics for day to day object, and therefore, we simplify our task by using classical physics for day to day world. But principally, quantum mechanics can very well be used for day-to-day motions. It has also been proven that Newton's second law of motion ($F = ma$) can be derived from Schrodinger's equation ($H\psi = E\psi$). Thus, there are no two worlds. Classical world is a special case or an approximation of quantum world

Before we conclude our discussion, we note that the quantum states described herewith are also called as stationary states, i.e. the states which do not change with time. Recall that we have approached this problem by using Schrödinger's time independent equation, as the potential energy was independent of time. These states (the energies and the wavefunctions) are also called as '*eigen*' states. The word '*eigen*' is a German word and it means '*characteristic*'

Example (9.1)

Calculate the first four energy levels of an electron trapped in a rigid box having width $1.0 \text{ } A^{\circ}$. Also calculate the first four energy levels of a marble of mass 10 gm trapped in a rigid box of width 1.0 m . Compare and interpret the results

Solution:

We have

$$E_n = \frac{n^2 h^2}{8mL^2}$$

For electron,

$$E_n = n^2 \left[\frac{(6.63 \times 10^{-34})^2}{8 \times 9.1 \times 10^{-31} \times (1 \times 10^{-10})^2} \right] J \times \frac{1}{1.6 \times 10^{-19}} \frac{eV}{J}$$

$$\Rightarrow E_n = (37.74) \times n^2 \text{ eV}$$

Thus, the first four energy levels are given in Fig (a)

For marble

$$E_n = n^2 \left[\frac{(6.63 \times 10^{-34})^2}{8 \times 10 \times 10^{-3} \times (1)^2} \right] J$$

$$\Rightarrow E_n = (5.5 \times 10^{-66}) \times n^2 J$$

These energy levels are shown in Fig (a). The difference between the consecutive energy levels is $\sim 10^{-66} \text{ J}$; too small to consider them discrete

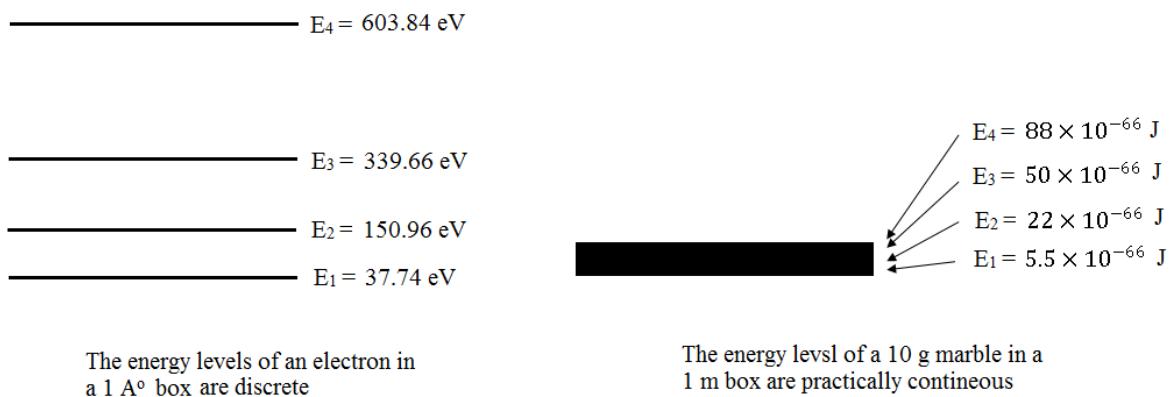


Figure (a) The energy levels of subatomic entities are discrete while those of day to day objects are continuous

Example (9.2):

Consider nucleus as a rigid box having width $\sim 10^{14} \text{ m}$, calculate the ground state energy of electron if it existed inside the nucleus and hence prove that electron cannot exist inside the nucleus

Solution:

We have

$$E_n = \frac{n^2 h^2}{8mL^2}$$

Considering ground state ($n = 1$) and taking $L = 10^{-14} \text{ m}$, we have

$$E_1 = \frac{1^2 \times (6.63 \times 10^{34})^2}{8 \times 9.1 \times 10^{-31} \times (10^{-14})^2} J \times \frac{1}{1.6 \times 10^{-19}} \frac{eV}{J} \times \frac{1}{10^6} \frac{MeV}{eV}$$

$$\Rightarrow E_1 = 3774 MeV \gg 8.8 MeV$$

As the maximum binding energy of the nucleus is $8.8 MeV$, the ground state energy of the electron if it existed inside the nucleus is too large to trap it there. Indeed the reality of the absence of the electron inside the nucleus is consistent with almost all principles of quantum mechanics and therefore it can be considered to be a test of validity of quantum ideas

Example (9.3):

What would have been the ground state energy of a cricket ball of mass $163 g$ in a room having width $10 m$, if were living in the world where Planck's constant was $6.63 J.s$? What would have been the minimum speed? Interpret the result.

Solution:

We have

$$E_n = \frac{n^2 h^2}{8mL^2}$$

$$\Rightarrow E_1 = E_{min} = \frac{(1)^2 \times (6.63)^2}{8 \times 163 \times 10^{-3} \times (10)^2} = 0.34 J$$

$$\frac{1}{2}mv_{min}^2 = 0.34$$

$$\Rightarrow \frac{1}{2} \times 163 \times 10^{-3} \times v_{min}^2 = 0.34$$

$$v_{min} = 2.03 \frac{m}{s}$$

It would not be possible to bring down the energy of the cricket ball below $2.03 \frac{m}{s}$. The cricket ball and indeed all the objects in such world would not be allowed to have a rest

Example (9.4):

Calculate the first four energy levels of the cricket ball in Example (9.3)

Solution:

We have

$$E_n = \frac{n^2 h^2}{8mL^2}$$

$$\Rightarrow E_n = (n)^2 \times \left(\frac{(6.63)^2}{8 \times 163 \times 10^{-3} \times (10)^2} \right) = 0.34 (n)^2 J$$

Thus the first four allowed energy levels are 0.34 J, 1.36 J, 3.06 J, 5.44 J. The corresponding allowed speeds are $22.03 \frac{m}{s}$, $4.09 \frac{m}{s}$, $6.13 \frac{m}{s}$ and $8.17 \frac{m}{s}$. A bowler, batsman or a fielder would not be allowed to throw or hit the cricket ball with any speed as per his choice. Or, you would not be able to accelerate your bike gradually and smoothly!

Example (9.5):

Calculate the quantum number associated with an electron moving in a rigid box of width 1 nm with a speed $10^6 \frac{m}{s}$ and a cricket ball of mass 163 g moving on a ground having size 138 m with speed $160 \frac{km}{h}$. Interpret the results.

Solution:

We have

$$\frac{1}{2}mv^2 = \frac{n^2 h^2}{8mL^2}$$

$$\Rightarrow n = \sqrt{\frac{8mL^2 \times mv^2}{2 \times h^2}}$$

$$n = \frac{2mLv}{h}$$

For electron

$$n = \frac{2 \times 9.1 \times 10^{-31} \times 1 \times 10^{-9} \times 10^6}{6.63 \times 10^{-34}} = 2.75 \approx 3$$

For cricket ball

$$n = \frac{2 \times 163 \times 10^{-3} \times 160 \times 10^3 \times \frac{1}{3600} \times 130}{6.63 \times 10^{-34}} \approx 3 \times 10^{36} !$$

It can be seen that, the quantum number associated with the electron is low, while that of a

cricket ball is extremely high. We recall that when quantum number acquires extremely high value, quantum physics gives the same results as of classical physics. Thus the subatomic particles have lower quantum numbers and they are quantum entities, but the entities in daily life are classical ones.

9.5 PARTICLE IN A NON-RIGID BOX (FINITE POTENTIAL WELL) (MATHEMATICS NEED NOT BE MUGGED UP, HOWEVER, THE CONCLUSIONS ARE IMPORTANT)

The wavefunctions penetrate the boundaries, even for insufficient energies

In previous section, we tackled the problem of particle in infinitely rigid box using Quantum mechanical treatments. Although, such infinitely box never exist, we approached this ‘ideal problem’, as that was the simplest one. The problem also allowed us to ‘practice’ quantum mechanics and learn its fundamentals. We now approach to a problem which is physically relevant...the particle in non-rigid box. We recall that, all motions in the nature, be it be motion of an electron in an atom or atom in a molecule, a proton or neutron in the nucleus or even a cricket ball on the ground, are the motions in non-rigid boxes. A thorough analysis of this problem will tell us about another set of quantum mechanical principles, which cannot be understood by our common sense based on daily experiences.

Finite potential well: An approach based on Heisenberg’s uncertainty principle.

Consider Fig. (9.5). An electron (as a wave) is incident on a potential barrier. The energy of electron (E) is less than the barrier height (V_0). Classically we expect the electron to bounce

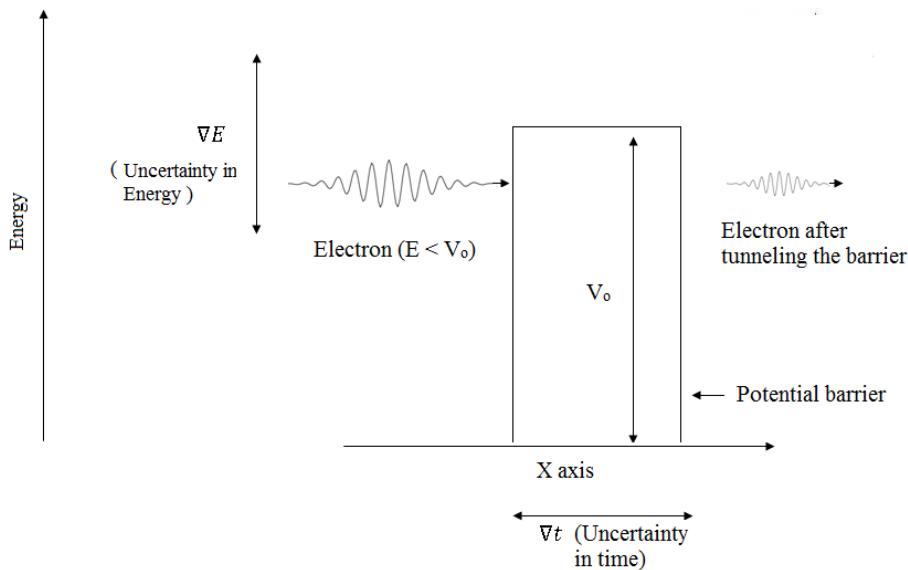


Figure (9.5) Barrier penetration (tunnel effect) explained using Heisenberg’s uncertainty principle

back, as its energy is insufficient to cross the barrier. However, the electron unexpectedly tunnels through the barrier. This can be very well explained by using Heisenberg's uncertainty principle. Assume that the electron has an energy E and correspondingly, a speed v . Assume that the width of the barrier is L . Thus if electron, if were allowed to cross the barrier, would take time $t = \frac{L}{v}$ to do so. Now according to Heisenberg's uncertainty principle, E and t can never be measured accurately. We have

$$\Delta E \Delta t \approx h$$

Thus the energy of the electron is uncertain by amount $\Delta E = \frac{h}{\Delta t}$. At the atomic scales, the barriers are sufficiently narrow and therefore can be crossed in lesser time. Thus t and consequently Δt are small enough. The correspondingly larger value of ΔE allows electron to possess an energy higher than the barrier height in time interval t . Thus the electron tunnels through the barrier. We may note that, before and after tunneling, the electron moves in an empty space, wide enough to demand extremely large time interval t to cover the space. Thus before and after the barrier, Δt is extremely large and thus ΔE is extremely small and thus in these regions, the electron's energy E is well defined and quite certain. We can frame this situation by saying that, when electron hits the barrier, it borrows an extra energy to cross the barrier in time t and safely returns it after tunneling through the barrier. This discussion also leads to a conclusion that an electron having energy less than, but close to barrier height will have a greater chance of tunneling. Similarly an electron will have a greater chance of tunneling the barrier, if it is thin. The ability of a subatomic particle to tunnel through a barrier without having an energy sufficient to do so is called as *tunnel effect* or *barrier penetration*. In forthcoming session, we will discuss it by using an approach based on Schrodinger's equation. But before we do so, we note that, tunnel effect, which may appear uncommon according to common sense (it is as if that a tennis ball hits on a hard wall and unexpectedly tunnels through it), it is used in nature and even in technology to a great extent. Indeed, two Nobel prizes in Physics, one in 1973 and another in 1986, have been awarded to Physicists who built ultra-sophisticated and technologically useful equipment using tunnel effect.

Tunnel effect: An approach based on Schrödinger's equation

Consider Fig (9.6) where a subatomic particle (say an electron) of mass m is trapped in a non-rigid box of length L . As the box is not infinitely rigid, the corresponding potential well is not infinitely deep. We thus treat it as a finite potential well having barrier height V_0 . Here, we carry forward some assumptions in the previous problem of infinitely rigid box, which are

1. Within the box, the particle is free, as the force acting on the particle is assumed to be zero. As the force is zero, the potential energy of the particle inside the box is zero
2. The motion is one dimensional
3. Even though the box has finite rigidness, the walls are infinitely thick

For mathematical convenience, we divide the problem in to three regions

1. Region I: where we have $-\infty < x < 0$ and $V = V_0$

2. Region II: where we have $0 < x < L$ and $V = 0$
3. Region III: where we have $L < x < +\infty$ and $V = V_0$

Recall that in previous problem, the particle with whatsoever energy was incapable of penetrating the walls of the infinitely rigid box. However, in this problem, because the barrier height V_0 is finite, the particle can have energy $E < V_0$ or $E > V_0$. The particle with $E < V_0$ is a bounded particle and therefore we are very much interested in analyzing its quantum mechanics. However, the particle with $E > V_0$ can easily cross the barrier and is therefore a free particle. We ignore this case, as there is no quantum mechanics for the free particle.

As such, the problem consists of regions I, II and III and therefore, although it is single particle moving in all regions, we intend to solve Schrodinger's equation for the same particle, in regions I, II, III. The solutions of Schrodinger's equation in these three regions, which can be represented by ψ_I, ψ_{II} and ψ_{III} , do not represent three different De Broglie waves, but three sections of the same de Broglie wave in three regions. It also appears that, as ψ_I, ψ_{II} and ψ_{III} represent the three sections of a single De Broglie wave, they ought to join smoothly at the boundaries, $X = 0$ and $X = L$. This indicates that these wave functions need to satisfy following boundary conditions (we cannot stay away from boundary conditions in quantum mechanics!).

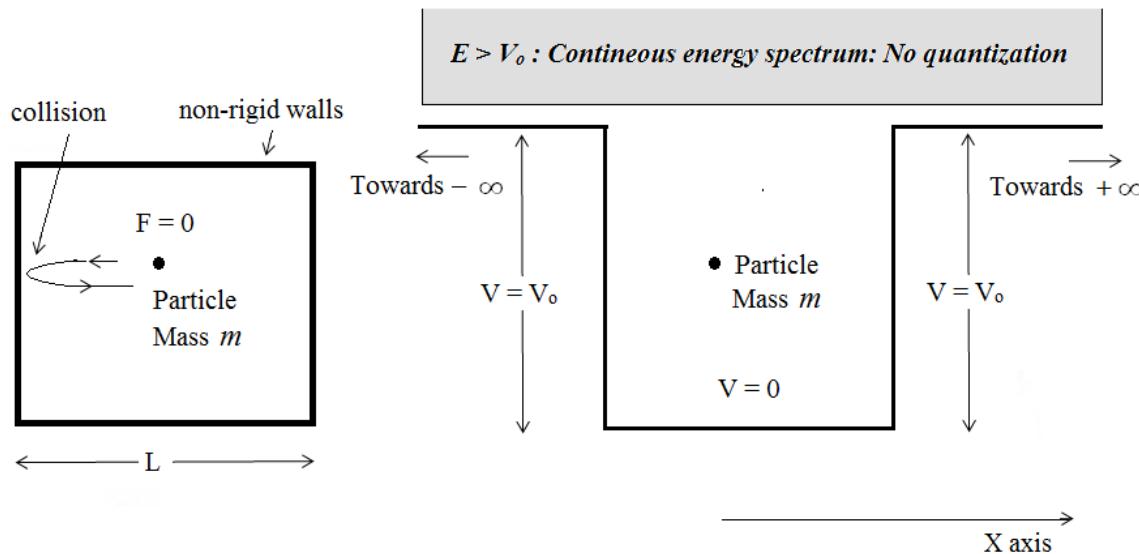


Figure (9.6): Particle in a non-rigid box

$$\psi_I = \psi_{II} \text{ at } X = 0, \frac{\partial \psi_I}{\partial x} = \frac{\partial \psi_{II}}{\partial x} \text{ at } X = 0, \psi_{II} = \psi_{III} \text{ at } X = L \text{ and } \frac{\partial \psi_{II}}{\partial x} = \frac{\partial \psi_{III}}{\partial x} \text{ at } X = L$$

Further, as the probability of finding the particle having finite energy at $X = \pm\infty$ needs to be zero, two more boundary conditions, as given below, need to be satisfied.

$$\psi_I \rightarrow 0 \text{ as } X \rightarrow -\infty \text{ and } \psi_{II} \rightarrow 0 \text{ as } X \rightarrow +\infty$$

A similar discussion in previous problem has already indicated that when we apply such boundary conditions to the wavefunctions, the energy is quantized. The above six boundary conditions cannot be satisfied for any arbitrary energy of the particle. They are satisfied only for specific values of the energies. Thus the energy levels for which boundary conditions are satisfied are the ‘allowed’ ones and those where the boundary conditions are not satisfied are ‘forbidden’. Thus we once again note that the energy of a ‘bounded (or trapped)’ particle (in finite as well as infinite potential wells) is quantized. We also recall that all motions in the nature are the motions with restrictions, the energy of any object in the nature, especially, the subatomic one, must be quantized. This itself is the bottom-line of ‘quantum’ mechanics.

We now proceed to attempt this problem using formal method of quantum mechanics; i.e. solving the Schrodinger’s equation. As the finite potential well in this problem is time independent, we solve this problem by using Schrodinger’s time independent equation. We have

$$-\frac{\hbar^2}{2m} \nabla^2 \psi + V(x, t)\psi = E\psi$$

As the motion is one dimensional, $\nabla^2 \psi$ reduces to $\frac{d^2\psi}{dx^2}$. We thus have

$$-\frac{\hbar^2}{2m} \frac{d^2\psi}{dx^2} + V(x, t)\psi = E\psi \quad \dots(9.45)$$

Rearranging,

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2} \{E - V(x, t)\}\psi = 0 \quad \dots(9.46)$$

By applying the above equation for regions I, II and III, we get

$$\text{Region I } (V = V_o), \quad \frac{d^2\psi_I}{dx^2} + \frac{2m}{\hbar^2} \{E - V_o\}\psi_I = 0 \quad \dots(9.47)$$

$$\text{Region II } (V = 0), \quad \frac{d^2\psi_{II}}{dx^2} + \frac{2mE}{\hbar^2} \psi_{II} = 0 \quad \dots(9.48)$$

$$\text{Region III } (V = V_o), \quad \frac{d^2\psi_{III}}{dx^2} + \frac{2m}{\hbar^2} \{E - V_o\}\psi_{III} = 0 \quad \dots(9.49)$$

As the quantities m , \hbar , E and V_o are constant, we introduce two constants, k^2 and k'^2 with following equations,

$$k^2 = \frac{2mE}{\hbar^2} \quad \text{and} \quad k'^2 = \frac{2m}{\hbar^2} \{V_o - E\} \quad \dots(9.50)$$

With these constants, the Schrödinger’s eqns (9.47), (9.48) and (9.49) turn in to

$$\text{Region I } (V = V_o), \quad \frac{d^2\psi_I}{dx^2} - k'^2\psi_I = 0 \quad \dots(9.51)$$

$$\text{Region II } (V = 0), \quad \frac{d^2\psi_{II}}{dx^2} + k^2\psi_{II} = 0 \quad \dots(9.52)$$

$$\text{Region III } (V = V_o), \quad \frac{d^2\psi_{III}}{dx^2} - k'^2\psi_{III} = 0 \quad \dots(9.53)$$

The solutions of Eqns. (9.51), (9.52) and (9.53) are

$$\text{Region I } (V = V_o), \quad \psi_I = Ae^{k'x} + Be^{-k'x} \quad \dots(9.54)$$

$$\text{Region II } (V = 0), \quad \psi_{II} = Pe^{ikx} + Qe^{-ikx} \quad \dots(9.55)$$

$$\text{Region III } (V = V_o), \quad \psi_{III} = Ce^{k'x} + De^{-k'x} \quad \dots(9.56)$$

Eqns (9.54), (9.55) and (9.56) for ψ_I, ψ_{II} and ψ_{III} contain six unknown constants, A, B, P, Q, C and D . These constants can very well be estimated by using six boundary conditions stated earlier

The condition $\psi_I \rightarrow 0$ as $X \rightarrow -\infty$ requires $B = 0$ in Eqn (9.54). Thus

$$\text{Region I } (V = V_o), \quad \psi_I = Ae^{k'x} \quad \dots(9.57)$$

The condition $\psi_{III} \rightarrow 0$ as $X \rightarrow +\infty$ requires $C = 0$ in Eqn (9.56). Thus

$$\text{Region III } (V = V_o), \quad \psi_{III} = De^{-k'x} \quad \dots(9.58)$$

Thus, out of six unknown constants, two are evaluated. These are $B = 0$ and $C = 0$

Applying condition $\psi_I = \psi_{II}$ at $X = 0$, we get from Eqn (9.54) and (9.55)

$$A = P + Q \quad \dots(9.59)$$

Applying condition $\psi_{II} = \psi_{III}$ at $X = L$, we get from Eqn (9.55) and (9.56)

$$Pe^{ikL} + Qe^{-ikL} = De^{-k'L} \quad \dots(9.60)$$

Applying condition $\frac{\partial\psi_I}{\partial x} = \frac{\partial\psi_{II}}{\partial x}$ at $X = 0$, we get from Eqn (9.54) and (9.55)

$$Ak' = Pik - Qik \quad \dots(9.61)$$

Applying condition $\frac{\partial\psi_{II}}{\partial x} = \frac{\partial\psi_{III}}{\partial x}$ at $X = L$, we get from Eqn. (9.55) and (9.56)

$$Pike^{ikL} - Qike^{-ikL} = -Dk'e^{-k'L} \quad \dots(9.62)$$

If we solve Eqns (9.59), (9.60), (9.61) and (9.62) as simultaneous equations, then the remaining constants, A, P, Q and D can also be evaluated. These evaluated constants can then be substituted in Eqns (9.55), (9.57) and (9.58). With this, we get completely known ψ_I, ψ_{II} and ψ_{III} . A careful look at these three wavefunctions indicates that ψ_I and ψ_{III} are exponential and ψ_{II} is oscillatory. We also note that, as X decreases from 0 onwards, ψ_I decreases exponentially, as X increases from L onwards, ψ_{III} decreases exponentially. Combining ψ_I, ψ_{II} and ψ_{III} for each allowed quantum state, we get the corresponding De Broglie's waves. These are shown in Fig (9.7). Note that, in each quantum state, the ψ_I, ψ_{II} and ψ_{III} join very smoothly. Recall that the quantum states, where ψ_I, ψ_{II} and ψ_{III} do not join smoothly are forbidden. Thus the energy is quantized.

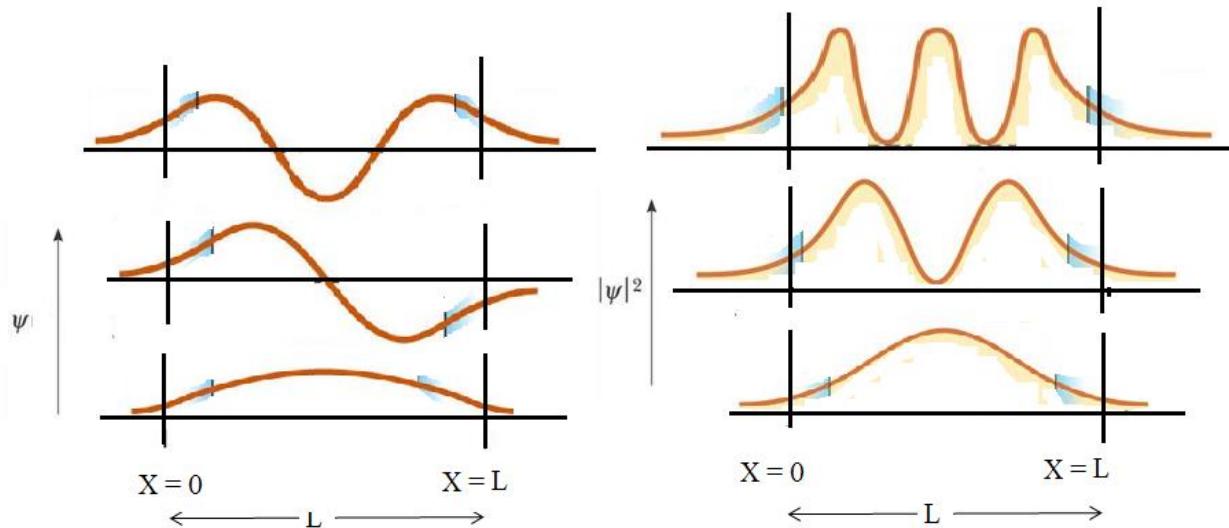


Figure (9.7): ψ and ψ^2 for finite potential well meant for first three quantum states

A careful look at these De Broglie waves indicates that, for each quantum state, the De Broglie waves are penetrating the classically forbidden regions. Recall that, we have solved this entire problem by considering $E < V_o$. Classically we expect that, a particle having its energy insufficient to cross a barrier can never cross it. However, the presence of ψ_I and ψ_{III} beyond $X = 0$ and $X = L$ indicates that a particle having an energy insufficient to cross a barrier can still cross it. In quantum mechanics, this phenomenon is called as *tunnel effect* or *barrier penetration*.

Let us have another look at this problem. We recall from Eqn. (9.50)

$$k'^2 = \frac{2m}{\hbar^2} \{V_o - E\} \quad \dots(9.63)$$

With an assumption $E < V_o$, we always have k'^2 to be positive and finite. Classically, the

condition, $E < V_0$ expects zero probability of finding the particle beyond $X = 0$ and $X = L$. This requires $|\psi_I|^2$ and $|\psi_{III}|^2$ to be zero beyond $X = 0$ and $X = L$. Consequently ψ_I as well as ψ_{III} need to be zero in these regions. But Eqns (9.57) and (9.58) indicate that $\psi_I = 0$ for $X < 0$ and $\psi_{III} = 0$ for $X > L$ requires k' to be ∞ , and this is never possible as long as $E < V_0$. Thus mathematics as well as Physics of this problem strongly support *tunnel effect (barrier penetration)*, where a particle having an energy lesser than the barrier height has a finite chance of tunneling the barrier

Before, we discuss the practical aspects of tunnel effect, we record two more observations

1. The wavefunctions penetrate to an increasingly greater extent in the barrier, for increasingly higher energy levels. But the probability of finding the particle decreases exponentially, as $X \rightarrow \pm\infty$.
2. The energy of a trapped particle is quantized
3. The wavelengths of the De Broglie waves in finite potential well are longer than those in infinite potential well. The relation $\lambda = \frac{h}{p}$ indicates that the momentum and consequently the energy levels the particles in a non-rigid box are lower than those in case of rigid box. This can be very well understood by recalling that the collisions of the particle with the rigid walls are perfectly elastic, while those with the non-rigid walls are inelastic.

9.6 TUNNEL EFFECT AND ITS APPLICATIONS

A particle having its energy insufficient to cross a barrier can still tunnel through it

Tunnel effect: Another look:

In previous section, the box was non-rigid, but the walls were infinitely thick, and therefore the electron, though tunneled in the wall, was permanently trapped there. If we make the walls (a barrier) thin enough, there are chances, although very low, of getting the electron on the other side. This situation is depicted in Fig (9.8). Instead of potential well, we now have a thin potential barrier of height V_0 and width L (small enough). Let an electron described by ψ_I be incident on the barrier. As discussed earlier, quantum mechanics allows this electron, even though having insufficient energy, to tunnel through the barrier. Let ψ_{II} be the wavefunction of the electron within the barrier and ψ_{III} be the wavefunction of transmitted electron. We may recall that in preceding section, the wavefunctions in open space were oscillatory and were exponential inside the barrier. All ψ_I , ψ_{II} as well as ψ_{III} can be obtained by usual procedure of quantum mechanics, *i.e.* is by solving Schrodinger's eqns, in the respective regions I, II and III. Note that, the amplitude (probability in this case) of ψ of the electron wave exponentially decreases inside the barrier and thus, the transmitted electron comes out with decreased amplitude *i. e.* probability (but same energy)

A rigorous mathematical analysis of this problem shows that

$$\text{Transmission probability(or transmission coefficient)} = T \approx e^{-2k'L} \quad \dots(9.64)$$

Where k' has same meaning as discussed in previous section, Eqn. (9.50)

$$k'^2 = \frac{2m}{\hbar^2} (V_0 - E) \Rightarrow k' = \sqrt{\frac{2m}{\hbar^2} (V_0 - E)} \quad \dots(9.65)$$

Eqns (9.64) and (9.65) collectively indicate that transmission probability decreases exponentially when width of the barrier increases or/and energy of the incident particle decreases. This characteristic of the tunnel effect has been incredibly used in some unique processes in the nature such as alpha decay, nuclear fission and fusion and even in technology such as tunnel diode, scanning tunneling microscope and SQUIDs (Josephson junctions). These inventions have led to Nobel prizes in Physics in 1973 and 1986.

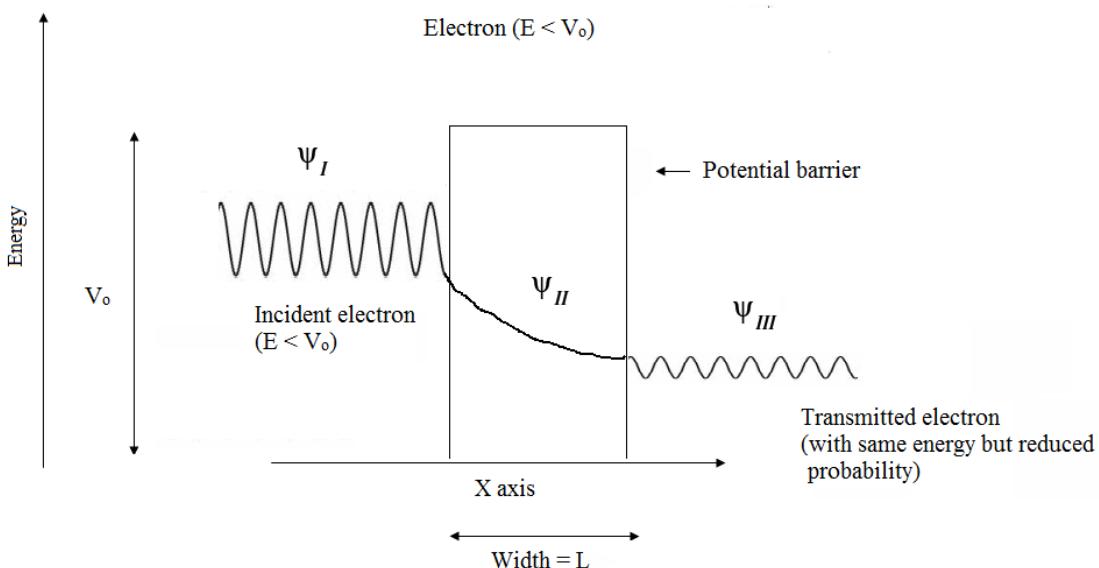


Figure (9.8): Tunnel effect: Another look

Practical aspects of Tunnel effect:

i. Alpha decay from radioactive nuclides:

We know that some heavy nuclides such as uranium or thorium emit alpha rays in the order to achieve stability. In nucleus, there is a very strong short range attractive force which keeps the nucleons together. This means that an alpha ray should overcome the attractive force inside the nucleus for coming out. Analysis shows that, typically, in the nuclides like uranium or thorium, alpha particles have to surmount a potential barrier of about 25 MeV for overcoming the inward attraction and thereby decaying out of the nucleus. Calculations indicate that these alpha particles possess energy in the range 4 to 9 MeV. This means that classically, alpha particles should never decay from these nuclides. But the experiments show that the nuclides such as uranium or thorium do emit alpha particles. What make this possible? Off course, it is tunnel effect. We

know that tunnel effect is very sensitive to the energy of the tunneling particles as well as the barrier width. In U^{238} , an alpha particle has to hit the barrier 10^{38} times, so that only once it gets a success in tunneling through the barrier. This amounts to 10^{21} times per second in the span of 4×10^9 years, which itself is the age of the earth!

Thus alpha decay clearly confirms tunnel effect. The theory of alpha decay based on tunnel effect was proposed by Gamow, Gurney and Condon in 1928. This was a striking confirmation of quantum mechanics through nuclear physics (interestingly Einstein's well-known $E = mc^2$ relation, which he proposed in 1915, was also verified in nuclear physics itself, when in 1919, the first nuclear reaction was discovered.

The rate at which alpha decay occurs is related with the half life of the element. The half-life of Th_{92}^{232} is 1.3×10^{10} years, which indicates slowest alpha decay, while the half-life of Po_{84}^{212} is 3.0×10^{-7} s! This indicates fastest alpha decay. This enormous difference is despite the fact that the energy of the alpha particle in Th_{92}^{232} (4.05 MeV) is only the half of its energy in Po_{84}^{212} (8.95 MeV). Also note that the mass numbers of Th_{92}^{232} and Po_{84}^{212} differ only by 20. This also confirms that the tunneling probability is quite sensitive to both, the energy of the particle as well as the height and width of the barrier.

As another example, consider the other two alpha particle emitters, U_{92}^{228} and U_{92}^{238} . The energy of the alpha particle in these two nuclides is 6.28 MeV and 4.25 MeV respectively. The energies differ only by 2.03 MeV. Further, the mass numbers of these two nuclides differ by only 10. However, the half life of U_{92}^{228} is only 9.1 min, while the half life of U_{92}^{238} is 4.5×10^9 years!. Tunnel effect is indeed very sensitive to energy of the tunneling particle and the height and width of the barrier!

ii. Nuclear fusion in sun and other stars:

We know that sun and indeed all the stars generate their enormous amount of energy by using nuclear fusion reactions. In such reactions, protons must fuse together to form deuterium and tritium nuclei. However, before fusing, the protons have to overcome Coulombic repulsion between them. To do so, they have to surmount an energy barrier of 1 MeV. The temperature of sun at its core is about 10^7 K. According to the relation $\left(\frac{1}{2}mv^2 = \frac{3}{2}kT\right)$ based on kinetic theory, the energy corresponding to this temperature is only **1 keV**, thousand times less than the required one. However, protons do fuse and therefore only sun is shining. What makes this possible? There are two reasons. The protons in sun follow Maxwell-Boltzmann distribution of energy, in which **1 keV** energy is the most probable energy at the peak. However, a few protons at the tail of the distribution have relatively higher energy, but even these protons have to use tunnel effect for penetrating the 1 MeV barrier. The tunnel effect, as it occurs here, is very weak. For the fusion of one proton pair, 10^{26} protons must collide with each other. The rate is thus hopelessly slow, but it is desirable, as otherwise sun would explode. Because of the huge mass of the sun, the available number of protons is so large that regardless of such a lower rate of fusion events in the sun's core, the deuterium is being produced at the rate of 10^{12} kg/s!

iii. Conduction of electrons through two copper wires twisted together:

Consider a copper wire cut in to two pieces. The pieces are then twisted together at the bare ends.

We know that when copper is exposed to atmosphere, its oxidation produces copper oxide which is an insulator. Thus, on the surface of the copper wire there is an insulating coating of copper oxide. However, in spite of this, the current still flows from one piece to another. This is due to tunnel effect.

iv. Frustrated total internal reflection (FTIR):

We very well know that when a light ray is incident on the denser to rarer interface at or greater than critical angle, it is totally internally reflected (Fig 9.9 a). However, this description of total internal reflection is based on ray optics which an approximation of wave optics. Wave optics proves that, even though a light wave is incident on a glass to air interface at a critical angle, instead of its total internal reflection, a small fraction of the beam is transmitted ahead in the second prism through a small air gap (barrier). (Fig 9.9 b). This is called as frustrated total internal reflection (FTIR) and it is another demonstration of tunnel effect. The air gap between the prisms is extremely small (a few optical wavelengths). Another explanation to FTIR is that, a light wave incident on the glass-air interface, it should go little-bit ahead to ‘know’ whether the next medium is air or not. FTIR finds applications in fiber optics

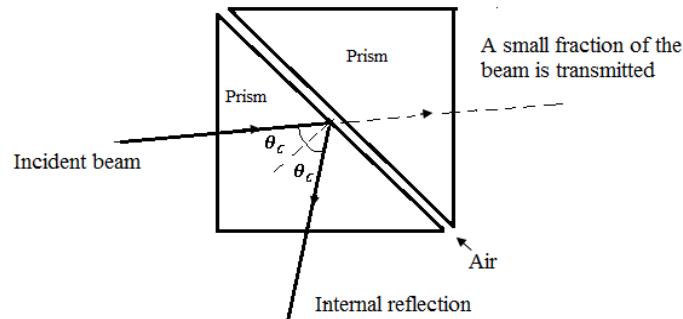
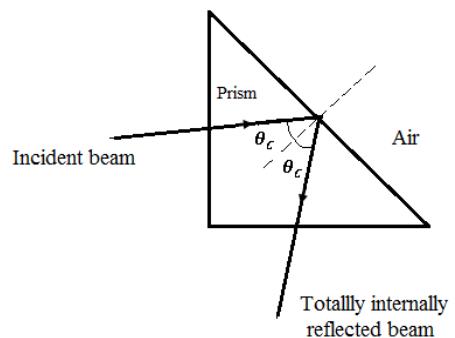


Figure (9.9): (a): Total internal reflection

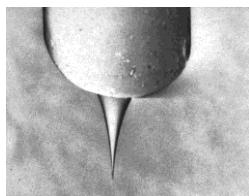
Figure (B): Frustrated total internal reflection (FTIR)

Applications of tunnel effect in technology

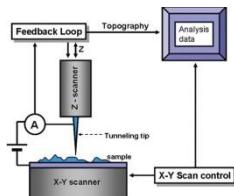
i. Scanning Tunneling Microscope (STM):

The main component STM is an extremely fine tip of tungsten (only one atom wide) whose motion can be finely and precisely controlled in X, Y, and Z direction by piezoelectric crystals. The tip is brought extremely close to the surface, whose atoms are to be imaged. A small potential difference (\sim millivolts) is applied between the tip and the surface to be imaged. In between the tip and the surface, there is a very thin air barrier. Regardless of the small potential difference, and the presence of air barrier in between the tip and the atoms, electrons from these atoms still tunnel through the barrier and enter in to the tip. This gives a tunnel current. As the tunnel effect is very sensitive to the barrier width, the tunnel current

decreases when the tip moves ahead and appears over the space between the atoms. Thus the data of the tunnel current obtained during scanning process can be used to produce a surface map of the atoms over the surface. When the STM is used in constant current mode, the position of tip is varied up and down to keep the tunnel current constant. Thus images of atoms can be produced. Note that STM is the only instrument which provides an image of an atom (recall that the atoms are million times smaller than the thickness of the human hair). The electrons tunneling from atoms to tip creates a loose contact between the atom and the tip. This helps in moving the atoms to the desired places. Nanoclusters with required shape and size can be made by manipulating the atoms in this way



(a) Extremely (one-atom-wide) tungsten tip in STM



(b) Schematic of STM



(c) STM facility in IISER, Pune

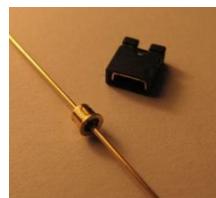


(d) An image "IBM" created by arranging the atoms using STM

Scanning Tunneling Microscope

ii. Tunnel diode

We know that the depletion zone in a PN junction acts like a barrier for the diffusion of majority charge carriers. This makes the diffusion process slow. Thus such diodes cannot be used in the circuits where faster response to change in voltage is required. However, by doping a diode heavily, its depletion zone can be made as narrow as $100 \mu\text{m}$. As the barrier is narrow, the majority carriers can tunnel through the barrier for even a low forward bias. The diode then responds rapidly to the changes in the voltage. Such diodes, which are rightly called tunnel diodes are therefore used in high frequency oscillators and as fast switches in the computers

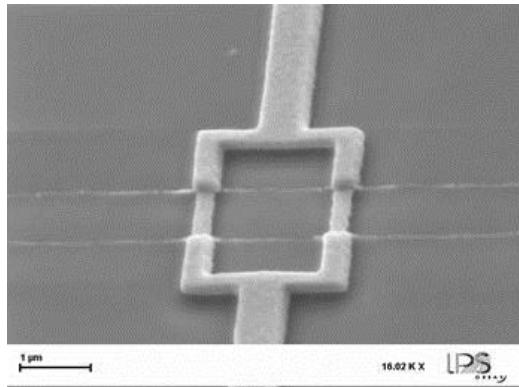


A tunnel diode

iii. Josephson junction:

Josephson junction contains two superconductors separated by an insulating barrier as thin as 2 nm. Due to thinness of the barrier, the wavefunctions of the Cooper pairs in the superconductors thus overlap and as a result, the Cooper pairs tunnel through the barrier. Josephson junctions are the part of ultrasensitive devices named as SQUIDs (Superconducting Quantum Interference Devices). If a SQUID is placed in the magnetic field, the current through the junctions varies very sensitively as per the changes in the magnetic flux through the loop. Extremely small

changes in magnetic fields, as small as 10^{-21} T, such as those in human body organs, for instance, brain stomach or even heart. SQUIDs are thus the ultrasensitive magnetometers, which can be used for medical diagnosis. Low-magnetic field MRI also uses SQUID magnetometers. The minute changes in the magnetic field of the earth before earthquake can also be detected by SQUIDs. SQUIDs can also measure the weak magnetic field produced by nanoparticles



A SQUID made up of two ultrathin Josephson junctions

9.8 QUANTUM MECHANICS AND ATOMIC STRUCTURE:

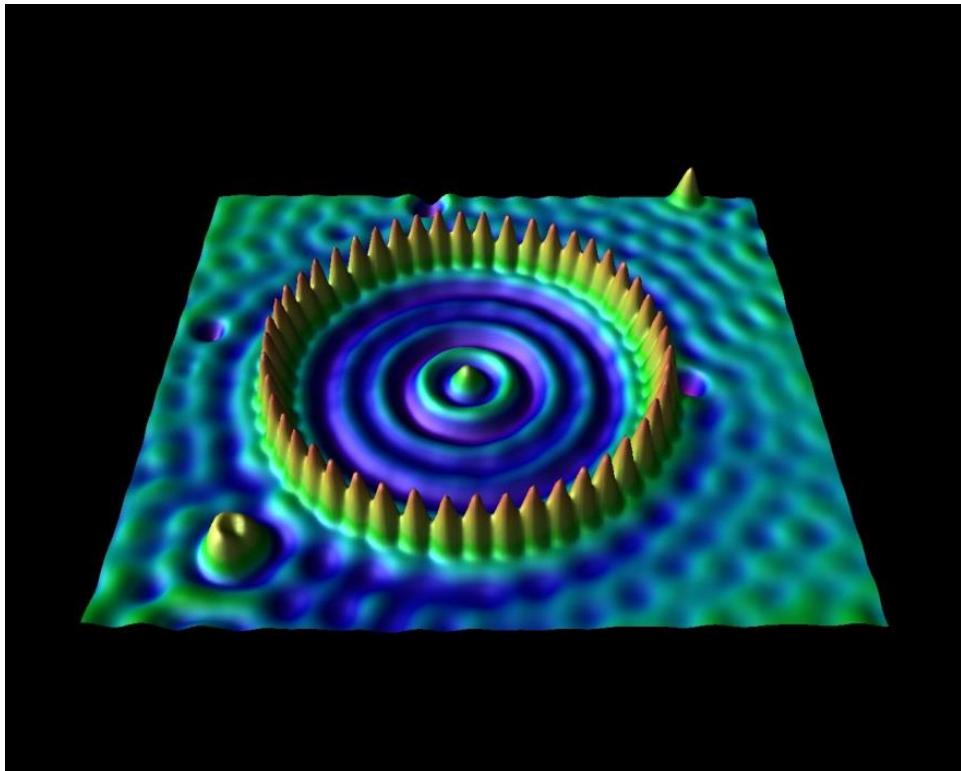
The quantum state of an electron in the atom is governed by four quantum numbers

An elaborate discussion of the two problems, namely infinite and finite potential wells has indicated presence of a quantum number ‘ n ’ as a result of restrictions on the motions. If we extend our discussion to three dimensional potential wells, then for each dimension (x, y and z), we require a separate wavefunction (ψ_x, ψ_y and ψ_z). As the motions are restricted in all three directions, there occur three quantum numbers n_x, n_y and n_z . The quantized energy can then be shown to be equal to

$$E_n = \frac{\hbar^2}{8mL^2} (n_x^2 + n_y^2 + n_z^2)$$

We now recall that an atom is a 3D box. Thus quantum mechanics of an atom should give rise to three quantum numbers. This problem was indeed taken up and then published by Erwin Schrodinger himself. As such, for an atom, the spherical polar coordinates (r, θ and ϕ) are more convenient. Consequently, the wavefunctions are (ψ_r, ψ_θ and ψ_ϕ). The three sets of boundary conditions for three dimensions (r, θ and ϕ) generate three quantum numbers namely (n, l and m_l). Out of these, n is called as *principle quantum number* and it decides the energy ($E_n = -\frac{13.6}{n^2} eV$) i.e. the *shell* of the electron in the atom. l is called as *orbital quantum number* and it decides the magnitude of angular momentum ($L = \sqrt{l(l+1)}\hbar$) i.e. the *subshell*, m_l is called as *magnetic quantum number* and it decides the orientation of the subshell

in the atomic space $L_Z = m_l \hbar$. A Nobel laureate Wolfgang Pauli introduced the fourth quantum number, named *spin quantum number* (m_s) to account for certain experimentally observed atomic properties. Thus in the Schrödinger's atomic picture, the quantum state of the electron in an atom is governed by four quantum numbers namely (n, l, m_s and m_s). Although , Schrödinger could solve the his equation only for hydrogen atom, its generalization to other atoms gave rise to rules of electronic configuration of even many-electron-atoms. Thus the atomic structure as we know it today is solely a creation of quantum mechanics, a strange theory of matter and energy at atomic scale. Before we conclude, we note that in 1913, Niels Bohr, a Nobel laureate, had thoughtfully introduced quantum ideas in atomic regime without knowing Schrodinger's equation (Bohr's quantum model of an atom came in 1913, while Schrodinger's equation was proposed in 1926). It has been proven that quantum mechanics does not straightaway discard the Bohr's model of the atom, which is based on a single quantum number, but extends it to four quantum numbers making it exceedingly accurate. One can appreciate the triumph of quantum mechanics in explaining the matter and energy at atomic details, if it is recalled that the properties and applications of energy and matter are principally governed by atomic structure.



In IBM Almaden Research Center in San Jose, California, the quantum corral was demonstrated in 1993 by Lutz, Eigler, and Crommie using an elliptical ring of 48 iron atoms on a copper surface using the tip of STM. The size of the corral is The De Broglie waves inside the nano traps can be easily seen. Matter exhibits wave-like properties at atomic scale

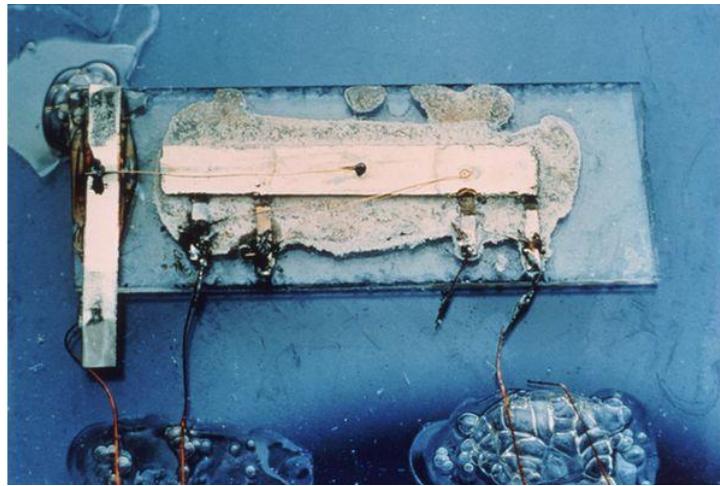
REFERENCE BOOKS

1. Fundamentals of Physics, 9th Edition, Extended, Wiley Resnick, Halliday, Walker,
2. Concepts of Modern Physics, Arthur Beiser, 6th Edition, Tata McGraw Hill
3. Introduction to Quantum Mechanics. - By D. Griffiths Published by Prentice Hall.
4. Quantum Mechanics. - By Ghatak and Loka Nathan Published by Mc. Millan
5. Quantum Mechanics. - By L. I. Schiff.
6. A Text-book of Quantum Mechanics by P.M. Mathews and K. Venkatesan, 2nd Edition, McGraw Hill Education, 2010
7. Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles, 2nd Edition, by Robert Eisberg , Robert Resnick, Wiley
8. Quantum Mechanics: Concepts and Applications, by Nouredine Zettili, 2016, Wiley

WORLD WIDE WEB

1. <https://quantumphysics.iop.org/>
2. <https://quantumphysicsmadesimple.com/>

Semiconductor Physics



The photograph on the left shows world's first transistor and on the right shows world's first integrated circuit. Both are Nobel Prize winning achievements. The old electronics was based on triodes and valves; and the corresponding circuits were bulky, slow and they required more power. The invention of transistor and integrated circuit overcame all these problems. The circuits based on transistor and ICs are compact, fast and require less power. The electronics thus has become efficient, compact and fast. Both these devices are based on semiconductors and their novel properties. What are semiconductors and how semiconducting devices work?

The answer to these questions is in this chapter

Contents

6.1 INTRODUCTION

Properties of semiconductors are the basis of today's electronics

6.2 ENERGY LEVELS OF INDIVIDUAL ATOMS:

They are discrete

6.2 FREE ELECTRON THEORY

A consistent explanation of electrical and thermal properties of metals

6.3 BAND THEORY OF SOLIDS

The energy levels of atoms become energy bands when a solid is formed

6.4 SEMICONDUCTORS

Moderate doping can enhance the properties of semiconductors to a great extent

6.5 FERMI DIRAC STATISTICS

Statistical description of electrons in semiconductors

6.6 FERMI LEVEL

A characteristic energy level which determines the behavior of semiconductors and semiconducting devices

6.7 INTRINSIC SEMICONDUCTORS:

Semiconductors in purest form, but without applications

6.8 EXTRINSIC SEMICONDUCTORS

How to enhance the conductivity of semiconductors

6.9 DESCRIBING SEMICONDUCTORS BY MEANS OF FERMI LEVEL

Fermi level of a semiconductor is governed by its type, temperature and doping level

6.10 QUANTITATIVE DESCRIPTION OF SEMICONDUCTORS:

How to characterize the semiconductors

6.11 HALL EFFECT

How semiconductors behave in magnetic field

6.12 PN JUNCTION DIODE ON THE BASIS OF ENERGY BAND DIGARM

Why diode conducts only in forward bias and not in reverse bias

6.13 TRANSISTOR ON THE BASIS OF ENERGY BAND DIGARM

How transistor amplifies the electrical signal

6.14 SOLAR CELL

A PN junction can convert sunlight into electricity

6.1 INTRODUCTION

Properties of semiconductors are the basis of today's electronics

X-ray diffraction has proven that crystalline solids have regular and periodic structures. Solid state physics deals with the properties and applications of electrical, optical, mechanical, thermal and magnetic materials. The properties and applications of any solid are governed by its two characteristics, one is its atom and another, the inter-atomic interaction. The electrical and electronic properties of solids show significant variations in different materials. Three distinct properties of copper, silicon and diamond are given in Table 6.1.

Table 6.2

Property ⇒ Material ↓	Type	Density of charge carriers n (m^{-3})	Resistivity ρ $\Omega\cdot\text{m}$	Temperature coefficient of resistivity (K^{-1})
Copper	Metal (conductor)	9×10^{28}	2×10^{-8}	$+4 \times 10^{-3}$
Silicon	Semiconductor	1×10^{16}	3×10^{-3}	-70×10^{-3}
Diamond	Insulator		10^{16}	

The table shows that there is a significant difference in the electrical properties of conductors, semiconductors and insulators. A quantum mechanical theory which explains this variation is called as band theory of solids. This theory also explains the behavior of many electronic devices such as diodes, BJT, FET, LED, Photodiodes, solar cells etc. Eleven elements in the periodic table show semiconducting properties. Out of these, silicon and germanium are widely used for manufacturing electronic devices. Germanium and silicon are called elemental semiconductors. Compound semiconductors such as gallium arsenide, cadmium sulfide, indium phosphide etc. are prepared by combining the elements from group III and V or groups II and VI. Amongst the several materials, the semiconducting materials have acquired a unique place in electronics due to their typical characteristics. These include

- i. The electrical properties can be significantly improved by adding impurities
- ii. Ability to conduct due to two kinds of charge carriers, namely electrons (negative) and holes (positive) (this makes the devices like PN junction diode, NPN or PNP transistor possible). Holes do not exist in most of the metals, and therefore, in spite of better conductivity, metals cannot be used to form diodes, transistors etc.
- iii. Temperature dependence of resistivity (this makes thermistors possible)
- iv. The band gaps are in the range of 1 to 3 eV. Due to this, semiconductors can convert light in to electricity and electricity in to light (this is the basis of optoelectronic devices such as LED, photodiode, solar cell, LDR etc.)

The invention of transistor in 1947 was a turning point in the history of technology. This invention proved to be a key in transforming old vacuum tube electronics to solid state electronics. The semiconducting devices such as PN junction diode, NPN or PNP transistor, field effect transistor, photodiode, LED, solar cells etc. have acquired a unique place in today's electronics. These devices offer several advantages over their vacuum tube analogs. These are

compact size, fast action, requirement of less power etc. This chapter aims at a thorough understanding of semiconductors and semiconducting devices. The discussion is based on band theory of solids and the concept of Fermi level.

6.2 ENERGY LEVELS OF INDIVIDUAL ATOMS:

They are discrete

The rules for electronic configuration of the atoms of various elements are solely governed by quantum mechanics. Accordingly, the electronic configuration of atom of any element is governed by four quantum numbers. These are principle quantum number (n), orbital quantum number (l), magnetic quantum number (m_l) and spin quantum number (m_s).

1. The principle quantum number (n) decides the energy that is shell of an electron. For $n = 1, 2, 3, 4$ etc, there are shells designated by K, L, M, N etc. n can take all integer values from 1 onwards. According to Quantum Mechanics, $n = 0$ is not permitted. The number of electrons accommodated in the shell are given by $2n^2$. There is no upper limit for n . Refer table 6.2

N	1	2	3	4	5	6
Shell	K	L	M	N	O	P
Capacity	2	8	18	32	50	72

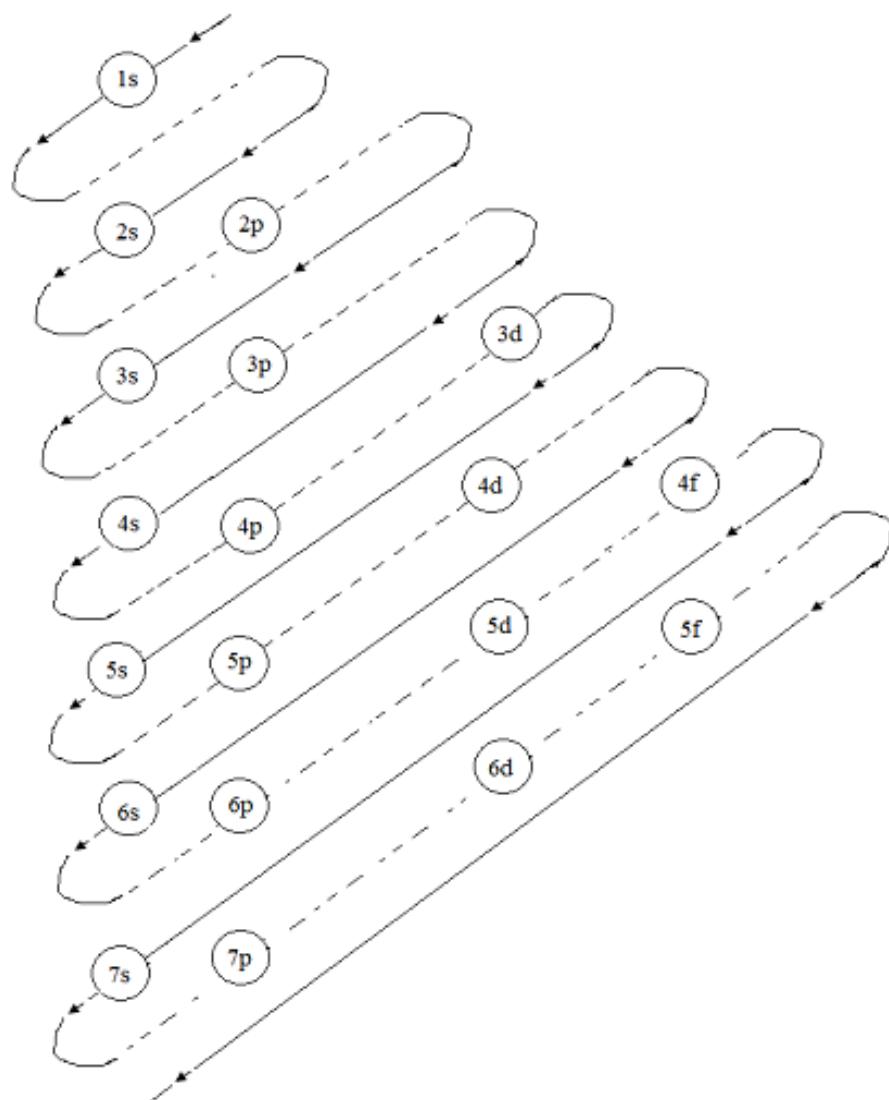
2. The orbital quantum number (l) decides the subshell of an electron within the shell. For a given value of n , the l can take values from 0 to $n - 1$. For $l = 0, 1, 2, 3$ etc., the corresponding subshell are designated by s, p, d, f etc. The total allowed values of l are n . The number of electrons in the subshell is given by n . Refer following table
3. The magnetic quantum number (m_l) decides the orientation of the subshell. For a given value of l , m_l can take values from $-l$ to $+l$ including 0. The total allowed values of m_l are $2l + 1$.
4. The spin quantum number (m_s) decides the spin of the electron. For every value of m_l , m_s takes two values which are $+\frac{1}{2}$ and $-\frac{1}{2}$.
5. Thus the occupancy of a subshell is given by $2 \times (2l + 1)$. The factor 2, outside the bracket accounts for the two allowed values of m_s , while $(2l + 1)$ accounts for $(2l + 1)$ allowed values of m_l .

l	0	1	2	3	4	5
Subshell	s	p	d	f	g	h
Capacity $2(2l + 1)$	2	6	10	14	18	22

The quantum state of an electron in an atom is governed by set of above four quantum numbers (n , l , m_l , m_s). Each quantum state corresponds to a different energy level. According to Pauli's exclusion principle, the quantum states of electrons in any atom are exclusively different. No two electrons can occupy same quantum state. The quantum states of electrons in the atom differ by at least one quantum number. This also means that once a particular quantum state (n , l , m_l , m_s) is occupied by a given electron, it is excluded for the other electrons.

The rules for electronic configuration of multi-electron atoms are as follows

1. **aufbau rule:** This rule states that, the electrons fill the subshells and shells with lower quantum number are filled at first. The corresponding electronic configuration acquires stability due to this rule. Following diagram gives a glimpse of aufbau rule.



- As discussed above, according to Pauli's exclusion principle, an orbital characterized by a given set of (n, l, m_l) can occupy only two electrons, having opposite spins $+\frac{1}{2}$ and $-\frac{1}{2}$
- Hund's rule:** According to this rule, pairing of electrons with opposite spins takes place only after an orbital has occupied the electrons with parallel spins.

All above rules signify that, any multi-electron atom prefers a configuration in which its energy is minimum and consequently has highest stability.

The electronic configurations of the elements which are involved in semiconductor physics (electronics), is given below

Element	Electronic configuration
Carbon (C, Z= 6)	$1s^2 2s^2 2p^2$
Silicon (Si, Z = 14)	$1s^2 2s^2 2p^6 3s^2 3p^2$
Germanium (Ge: Z= 32)	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^2$
Tin (Sn: Z = 50)	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^2$

Table 6.3 Electronic configuration of group IV elements

It can be noted that all these elements are tetravalent. As we shall notice later, diamond is insulator, silicon and germanium are semiconductors and tin is a conductor.

6.3 FREE ELECTRON THEORY (OPTIONAL)

A consistent explanation of electrical and thermal properties of metals

We know that metals and their alloys such as silver, copper, aluminum, lead behave like conductors. Their electrical and even thermal properties are superior as compared to the other materials. This behavior is explained by classical free electron theory, which was developed by Paul Drude in 1900, just three years after the discovery of the electron by J.J. Thomson. This theory was later refined by Hendrik Lorentz in 1905. Consequently, this theory is also called as Drude-Lorentz theory. Some limitations of theory were overcome by quantum free electron theory, developed by Sommerfeld in 1928. Further, the limitations of even quantum free electron theory were overcome by band theory of solids developed by Felix Bloch

Concept of a free electron (fundamental notion of the classical free electron theory):

The fundamental assumption of the classical free electron theory is that, in metals (monovalent, divalent and trivalent), the valence electrons from the atoms become free when the metals are formed. This is because, in metals, the atoms are bound together by weak metallic bonds. Further the energy of an electron belonging to a complete metal is lower than it would be if the electron were with the individual atoms. As the valence electrons become free, the atoms thus left behind are called positive ion cores. These ion cores are immobile, and suffer only minute thermal vibrations around their mean positions. These immobile ion cores are situated across the lattice points in the periodic lattice. At the room temperature and in absence of external electric field, the free electrons move randomly in all the directions. In fact, such electrons are free to move throughout the metal. During this motion, they suffer collisions due to each other and also due to

positive ion cores embedded within the metal. This random motion depends upon the temperature, and therefore called as thermal motion. This motion is similar to the motions of atoms or molecules in a gas and therefore the cloud of such free electrons is also called as ***free electron gas***. The electrical and thermal properties of metals are governed by such free electron gas. The energy related to the vibrations of the ions at the lattice points is quantized and is called as a phonon. During each collision with the ion, electron changes the direction. Due to this, the net velocity of the electron in the metal is zero. Such velocity is called as ***thermal velocity***. The average path covered by the electron during the successive collisions is called as ***mean free path***. The mean free path of the electrons in metals is few A° . The sea of such free electrons in metals is responsible for their electrical, thermal, optical and magnetic properties. For ex. the opacity, luster, high electrical and thermal conductivities of the metals can be explained using this classical free electron theory. The ability of the metals to combine with each other can also be explained using this theory.

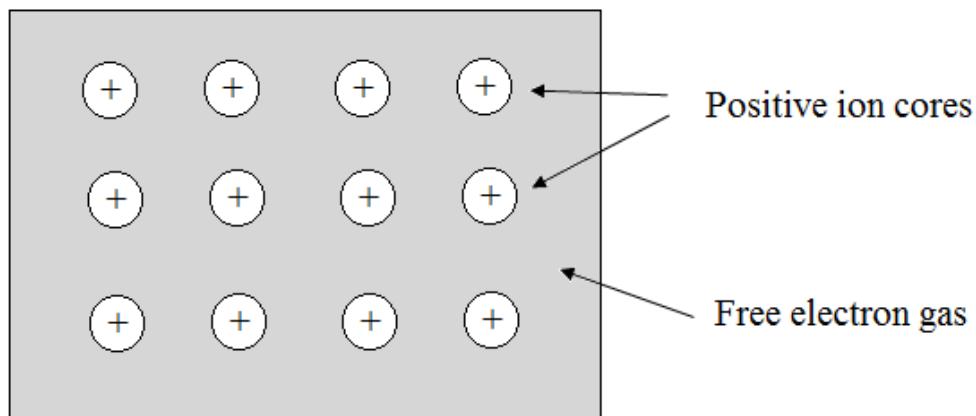


Figure 6.1The model of a metal according to classical free electron theory

Salient features of the classical free electron theory:

1. In metals, the valance electrons are free to move and hence they are the conduction electrons. In such free electron gas, there are positive ion cores distributed symmetrically. Normally, the ion cores are from all the sides of an electron. Thus the resultant forces of attraction by the ion cores on the electrons are zero. Consequently the potential inside the metals is due to the positive ion cores and it is uniform, throughout the metal. The electrons cannot escape the metal as at the boundary, the potential changes suddenly and becomes high. The electrons are thus trapped in a ‘metallic box’ between the high potential barriers at the boundaries.
2. The electrostatic forces of attraction between the electrons and positive ion cores and the electrostatic forces of repulsion between the electrons are negligible.
3. **Thermal motion of electrons in the metals:** According to classical theory the energies of the electrons in metals are distributed according to the classical Maxwell-Boltzman

statistics (which is mainly applicable to the molecules in the gas). According to this statistics, the kinetic energy of the electrons in motion is given by

$$\frac{1}{2}mv^2 = \frac{3}{2}kT \quad \dots(6.1)$$

Thus the kinetic energy solely depends upon the temperature. The corresponding motion of these electrons is thus called thermal motion. If we substitute $T = 300$ K, then the corresponding kinetic energy of the electrons turns out to be approximately 0.025 eV. The corresponding thermal speed is roughly 10^6 m/s. During this motion, the electrons suffer collisions with themselves and with the positive ion cores. As a result, in each collision, the direction of the electrons changes randomly. Due to such random collisions, the electrons move in virtually all possible directions and in zigzag paths. As a result the net resultant velocity is zero. Thus thermal motion does not give electrical current.

4. **Drift motion of electrons in metals:** When electric field is applied, the electrons start drifting in a direction opposite to the electric field. Here too, the electrons collide with themselves and with the ions in their path. However, the motion is not absolutely random, on an average the electrons drift towards the positive potential. The drift velocities of the electrons in the metals are roughly 10^{-2} m/s, which are very less as compared to thermal speed. The drift motion results in to ***drift current***.

Metallic Hydrogen

It has been realized that, hydrogen, if subjected to extremely high pressure, exhibits metallic properties. Such metallic hydrogen is expected to have applications in superconductivity, efficient rocket propellant, and nuclear fusion. Metallic hydrogen exists in the planet Jupiter and is responsible for Jupiter's high magnetic field

Drawbacks of classical free electron theory.

Despite that fact that classical free electron theory explained the high electrical and thermal conductivities, Ohm's law and a few other properties of metals, this theory has some drawbacks which are discussed below

1. This theory proposes that the electrical conductivity depends upon the concentration of the free electrons. This requires that trivalent(aluminium, indium etc.) and divalent metals (cadmium and zinc) should have better conductivity than the monovalent metals (copper, silver etc.) However, experimentally it is found that copper and silver have better conductivity than divalent and trivalent metals.
2. This theory proposes that the conductivity of the metals is due to negatively charged electrons. Thus all the metals should have negative Hall coefficient ($R_H = \frac{1}{nq}$). However some metals such as zinc have positive Hall coefficient. This cannot be explained by using classical free electron theory.
3. The average distance covered by the electron between two successive collisions is called as ***mean free path***. Classical free electron theory predicts that the mean free path should

be of the order of 3 \AA^0 . However, it is experientially found that typical mean free paths of the electrons in the metals are of the order of 50 \AA^0 .

4. Classical free electron theory predicts that the resistivity of the metals should be proportional to the square root of the absolute temperature. However experimentally it is found that the resistivity of the metals is directly proportional to the absolute temperature.
5. Classical free electron theory predicts that the molar specific heat at constant volume is given by

$$C_V = \frac{3}{2}R$$

However, experimentally it is found that,

$$C_V = 10^{-4}RT$$

Thus experimentally C_V is much smaller and also temperature dependent. This is against classical theory

6. The classical free electron theory predicts that the magnetic susceptibility of metals to be inversely proportional to the temperature, however experiments indicate that it is independent of the temperature.
7. This theory fails to explain the classification of the solids in to conductors, semiconductors and insulators.

To overcome these drawbacks, a quantum free electron theory was developed by Somerfield in 1928.

Quantum free electron theory

This theory retains the following vital features of the classical free electron theory

1. The metal consists of the free electron gas in which the positive ion cores are symmetrically distributed
2. Both the electrostatic force of attraction between the electrons and positive ion cores and the electrostatic force of the repulsion between the electrons are negligible.
3. The electrons in the metals are trapped in the metallic potential well having constant potential. The well is surrounded by the high potential barrier at the boundaries.

The fundamental difference between the classical and quantum free electron theory is related to the energy distribution of the electrons. As seen earlier, the classical free electron theory suggests that the energies of the free electrons in the metals are distributed according to Maxwell Boltzman statistics. However, quantum theory proposes that the electrons are distributed in the various energy levels according to Fermi Dirac statistics which is mainly based on Pauli's exclusion principle. The Pauli's exclusion principle suggests that a given quantum state can be occupied by only one electron. Thus in a set of continuously distributed energy levels, every level occupies two electrons. The electrons fill up the levels in the increasing order of energy. The highest occupied energy level at 0 K is called as Fermi level (E_F). In metals, the kinetic energy of the free electrons is considerably higher as compared to their kinetic energy.

Limitations of Quantum free electron theory:

The quantum free electron theory overcomes certain drawbacks of the classical theory; however, even this theory has some drawbacks, one of which is that it cannot explain the classification of the solids into conductors, semiconductors and insulators. The quantum theory also fails to explain the positive Hall coefficients of some elements like zinc. This drawback was overcome by **band theory of solids** which was developed by Felix Bloch around 1928. It may be noted that this theory is rooted in quantum mechanics, and indicates that quantum mechanics works for individual as well as aggregate of atoms.



Felix Bloch (1905-1983) He was a Swiss Physicist, who obtained degree in engineering in Switzerland. He did his Ph.D. in Physics in Germany under Werner Heisenberg. In his Ph.D. thesis, he showed the origin of allowed and forbidden energy bands in solids by solving the Schrödinger's equation for electron in periodic crystal. He worked as a Professor of Physics in Stanford University (US). During his career, he was closely associated with Heisenberg, Schrödinger, Pauli, Bohr, and Fermi, all of whom are Nobel laureates. During 1954-55, he became the first Director-General of CERN. He, along with Edward Purcell developed the theory of Nuclear Magnetic Resonance (NMR) for which he received Nobel prize in Physics in 1952. NMR is the basis of Magnetic Resonance Imaging (MRI), a Nobel prize winning medical diagnostic technique. He was also involved in the development of the first atomic bomb.

6.4 BAND THEORY OF SOLIDS

The energy levels of atoms become energy bands when a solid is formed

The electrical properties show a drastic variation amongst the various solids. For ex. the conductivity of copper is 10^{25} times greater than diamond. A typical theory which explains this behavior is called as band theory of solids. This theory can be arrived at by two approaches, one developed by Felix Bloch (restrictions on the allowed energy levels of the electrons due to the periodicity of the crystal) and another by Walter Heitler and Fritz London (splitting of the energy levels of the electrons due to their interaction). We will follow the later.

Consider an atom of copper, which contains 29 electrons. According to Pauli's exclusion principle, these electrons are distributed in to 29 exclusively different quantum states characterized by a group of four quantum numbers (n, l, m_l, m_s). Out of these, the first two quantum numbers (n, l) have a stronger impact on the energy levels of the electrons while the impact of (m_l, m_s) is insignificant. The electronic configuration of copper is as follows

$$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^1$$

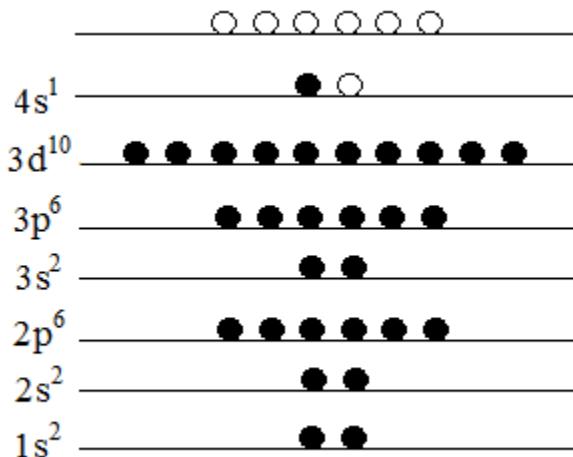


Figure (6.2): Energy levels in an atom of copper

The corresponding energy levels of the electrons are as shown in the Fig (6.2). The energy levels of the atom of copper (or any other element) are always discrete. Now if we consider two copper atoms at infinite distance, then such atoms do not interact with each other in any way. In such case the $29 \times 2 = 58$ electrons are distributed in the same energy levels, as in the first copper atom. If there are N number of copper atoms, each at an infinite distance from another, then we can say that, 1s **level** will accommodate 2N electrons, 2s (2N electrons), 2p (6N electrons), 3s (2N electrons), 3p (6N electrons), 3d (10N electrons) and 4s (N electrons). Now, consider that two copper atoms, initially at infinite distance are gradually brought closer to each other. When the atoms come sufficiently close, initially the outermost that is valance electrons will start ‘feeling’ each other’s presence. At equilibrium, the inter-atomic distance of the copper atoms in the solid form is 2.6 \AA° . If the atoms come close to each other at such distance, the valance electrons will start strongly interacting with each other. In such case the electrons cannot be accommodated in the same energy levels. For ex. the two 4s electrons cannot be accommodated in the same 4s level. The 4s level splits in to two closely spaced 4s sublevels. This is because,

- i. The electrons undergo a strong electrostatic interaction hence their energies change.
- ii. According to Quantum Mechanics, the electrons behave like waves. Thus when electrons come in the proximity, their wavefunctions overlap. As a result, symmetric and anti-symmetric wavefunctions, which correspond to two distinct energy levels, are formed.
- iii. When the two atoms are at infinite distance, they are two different systems, and thus Pauli’s exclusion principle does not apply to such pair of copper atoms. However, when the atoms come close and form bonds, a pair of two atoms becomes a single system and thus Pauli’s exclusion principle applies to the pair. Accordingly, there have to be two different quantum states with different energies for the two 4s electrons. With the same logic, when three copper atoms come together, the 4s level splits in to three different energy levels, each accommodating a different 4s electron. Thus when N atoms come together, there will be N number of energy levels, all belonging to 4s category. Typically 1 cm^3 of a solid consists of roughly 10^{23} atoms.

When these atoms form a solid, there occur 10^{23} distinct levels all belonging to 4s category. As a result, these split energy levels are infinitesimally close to each other with a separation $\sim 10^{-23}$ eV. The levels are thus virtually continuous. A group of such 4s levels thus becomes a 4s band. (Refer

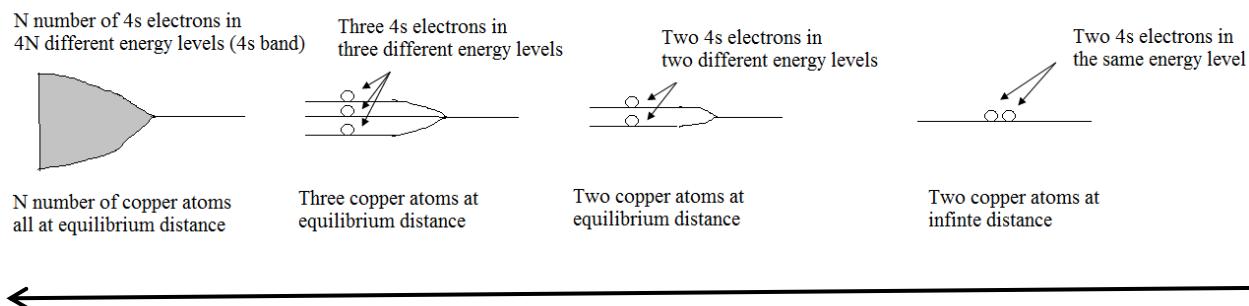


Figure (6.3): Transformation of energy levels in to energy bands

Similar effect is observed for rest of the energy levels. The wavefunctions of these electrons also overlap and the levels split and form bands. Refer Fig (6.3). One can note that, in each band, there are as many energy levels as there are atoms (N). The number of quantum states in the band is the product of occupancy of the corresponding level and the number of atoms. Thus s band consists of $2N$ states, p band contains $6N$ states, d band contains $10N$ states and so on. It may be noted that originally, in the energy level diagram of an atom, the energy levels are discrete and in between any two allowed energy levels, there are a few forbidden energy levels. This trend continues for energy bands also. Thus in between any two allowed energy bands, there are forbidden bands, which consist of forbidden energy levels. Thus, the energy levels for individual atoms become energy bands when atoms are aggregated. An atom is characterized by its own energy level diagram, while the corresponding solid (the aggregate of atoms) is characterized by energy band diagram. The energy band diagram is characterized by the total number of bands which it involves, their occupancy, size and the forbidden gap between them. The energy band diagrams of different solids are different, as different solids have different types of atoms separated by different distances. As we shall see later, some energy band diagrams have two distinct features, namely, the partially or completely filled bands and overlapping of consecutive bands. The electrical as well as some other properties of the solids are solely governed by their energy band diagrams. Fig.(6.4) shows the typical energy band diagram of copper. It can be noted that 4s band has an occupancy of $2N$ states, however, it contains only N number of electrons occupying the lower levels. 4s band is thus partially (half) filled. As we shall see later, such partially filled band imparts conducting properties to the solids.

Fig....also indicates that, the higher band have greater width as compared to the lower bands. Indeed the bands become narrower, when we move down. This is because; the upper bands are due to stronger interaction of valence electrons. As we move down, the lower level electrons interact weakly due to greater distance and screening effect of the upper level electrons. Due to larger distance, the wavefunctions of these electrons overlap to a lesser extent. As a result the splitting effect is weak and therefore the bands are narrower.

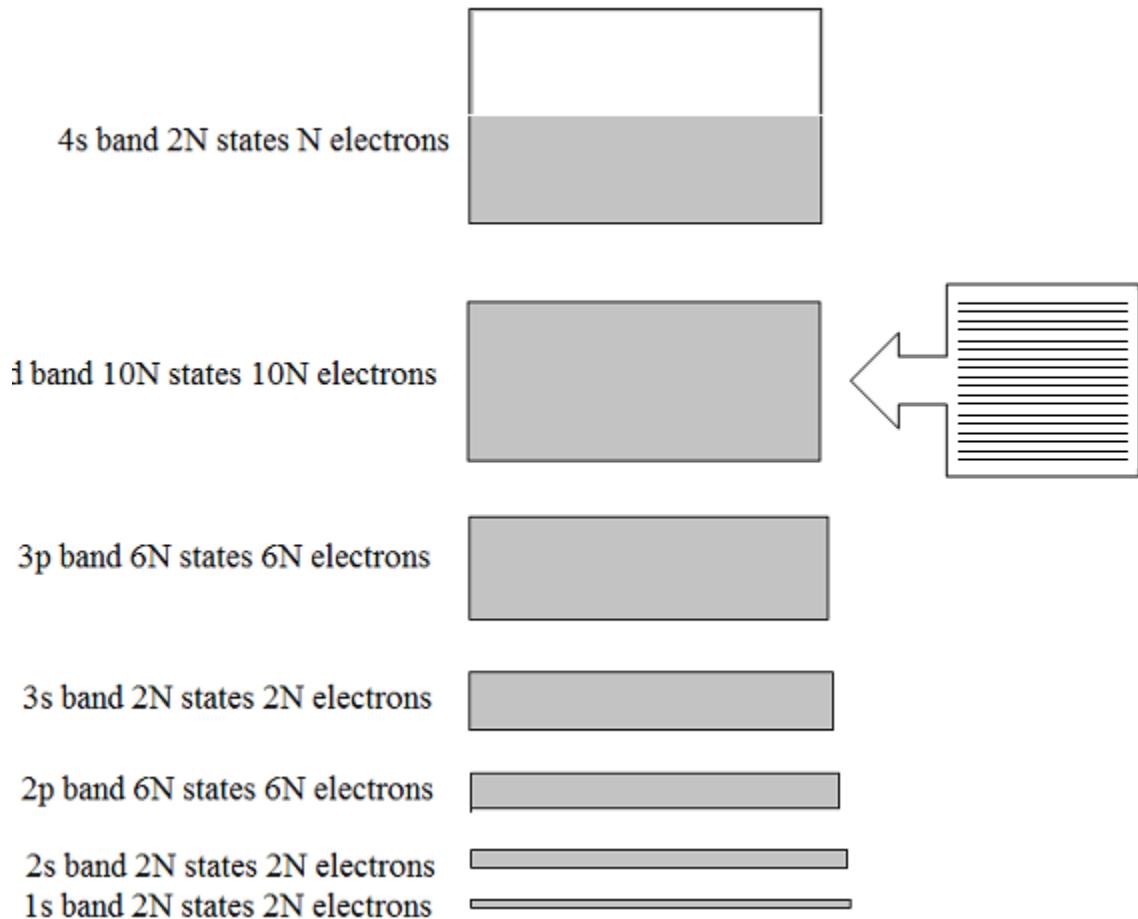


Figure (6.4): The energy band diagram of solid of copper

Valance band, conduction band and forbidden gap:

The electrons in the outermost orbit of any element are called as valance electrons. Consequently, the group of energy levels of the valance electrons constitutes a ***valance band***. Alternatively, valance band is also defined as the topmost or highest occupied band. All the bands above the valance bands are empty. In the group of unoccupied bands, the lowest unoccupied band is called as ***conduction band***. When the electrons in the valance band are bound to their parent atoms, but when these electrons are excited to conduction band, they become free and thus take part in conduction. In between the valance band and conduction band, there is a group of forbidden energy levels. Quantum mechanically these levels are not allowed to the electrons. This group is called as a forbidden gap or the band gap. As we shall discuss it further, the electrical properties of all solids are exclusively governed by the width of the forbidden gap.

Let us now discuss the energy band diagrams of some specific elements

Lithium ($Z = 3$)

The electronic configuration of lithium is

$$1s^2 2s^1$$

As shown in Fig (6.5), the energy level diagram consists of 1s level having two electrons and 2s level having 1 electron. As discussed earlier, when lithium atoms come together and form bonds, the energy levels split and form energy bands.

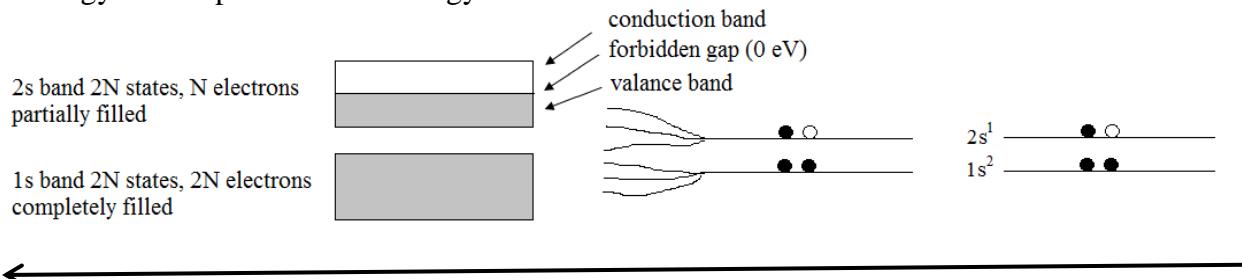


Figure (6.5): Band formation in lithium

Thus for lithium solid, there are two bands, namely 1s band and 2s band. 1s band has $2N$ states filled by $2N$ electrons. Thus 1s band is completely filled. However, the case is not so for 2s band. 2s band has $2N$ available states, but as the 2s level for lithium atom contains only 1 electron, the 2s band contains only N number of electrons. According to auf-bau rule, these electrons fill up the states in the increasing order of energy. Thus the lower N states are filled with N electrons and upper N states remain unoccupied. Thus 2s band of lithium is partially (half) filled. The lower filled part itself is the highest occupied band that is valence band, while in the same 2s band, the upper unfilled band is the lowest unoccupied band that is conduction band. As can be seen from fig (6.5), the gap between these two bands is almost 0 eV (actually 10^{-28} eV). According to kinetic theory, the thermal energy of the electrons is given by $\frac{3}{2}kT$. At 0K, there is no energy available for the electrons to rise in the conduction band, however, at 300 K, the thermal energy is 0.038 eV, which is sufficient for the electrons to go in the conduction band. It may be noted that there are no forbidden levels in between the valence and the conduction band. The allowed but empty energy levels in the conduction band are in the vicinity of the filled level in the valence band. When the electrons are excited to the conduction band, they become free. As the valence electrons can be made free very easily, lithium, at room temperature, has plenty of electrons in conduction band, and consequently, lithium behaves as a conductor. Lithium belongs to group I in the periodic table. The other elements in group I, such as sodium ($Z=11$), potassium etc. also have partially filled valence bands and thus are conductors. This indicates that any element having partially filled valence band should behave like a conductor. The other examples of the elements with partially filled valence band are, aluminium ($Z=13$), copper ($Z=29$), silver ($Z=47$) and gold ($Z=79$). As we know, all these elements are good conductors of electricity.

Beryllium ($Z = 4$).

The electronic configuration of beryllium is as follows

$$1s^2 2s^2 2p^{\dots}$$

Based on the discussion till now, we expect that the 2s band of the beryllium should be completely filled ($2N$ states filled with $2N$ electrons) and should behave like the valence band. The next band should be 2p, it should be empty and should behave like a conduction band. We expect that because of the existence of a finite gap (containing forbidden levels) between the valence and conduction band, beryllium should behave like an insulator. However beryllium also behaves like a conductor. This is because, when beryllium atoms come close to each other, the 2s and 2p bands split. As the distance approaches equilibrium separation, the 2s and 2p band expand and at certain stage the merge (overlap) in to each other. Thus instead of separate 2s and 2p bands, we get 2s-2p hybrid band having $2N + 6N = 8N$ available states. Out of these $8N$ states only lower $2N$ states are filled while the upper $6N$ states are unfilled. Thus 2s-2p hybrid band is partially filled and thus beryllium behaves as a conductor. Beryllium belongs to group II elements in the periodic table. The other elements in the periodic table such as magnesium also have overlapping of bands and are thus the conductors of the electricity

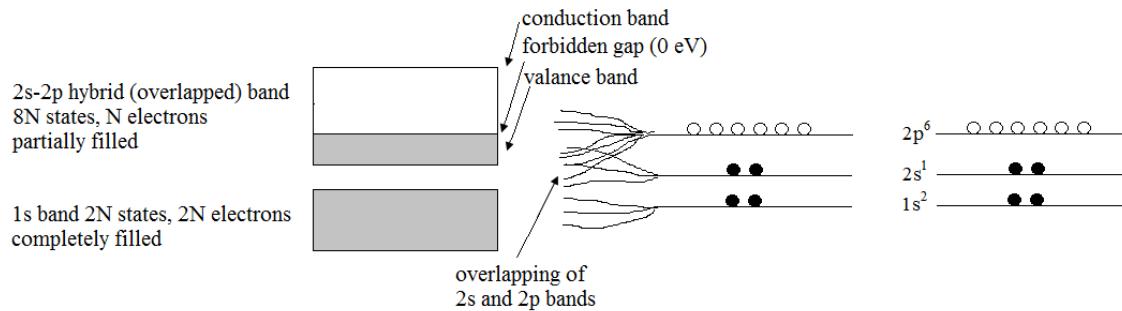


Figure (6.6): Energy band formation in beryllium

Similar effect occurs in magnesium due to overlapping of 3s and 3p bands, and as a result, it is conductor. Overlapping (hybridization) s and p bands also occurs in diamond ($Z = 6$), silicon ($Z = 14$), germanium ($Z = 32$) and tin ($Z = 50$). However diamond behaves as an insulator, silicon and germanium behave as semiconductors and tin behaves as a conductor. Why? We will discuss this in the next section.

Energy band formation in group IV elements:

Table 6.4 enumerates the properties of group IV elements. It can be noticed that the outermost orbit of all these elements contains 4 electrons. These elements are thus tetravalent. Each atom is surrounded by four atoms. Thus each atom forms four covalent bonds with four neighbors. In each bond the pair atoms share their two electrons.

Now consider Fig (6.8) . When the inter-atomic distance of the diamond atoms is very large, the energy levels 1s, 2s and 2p accommodate $2N$, $2N$ and $2N$ electrons respectively.

Sr. No.	Element	Symbol	Atomic Number (Z)	Electronic configuration	Inter-atomic separation (A°)	Band Gap (eV)
1	Diamond	C	6	$1s^2 2s^2 2p^2$	3.56	5.7
2	Silicon	Si	14	$1s^2 2s^2 2p^6 3s^2 3p^2$	5.43	1.12
3	Germanium	Ge	32	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^2$	5.66	0.72
4	Tin	Sn	50	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^2$	6.46	0.08

Table 6.4 Properties of group IV elements

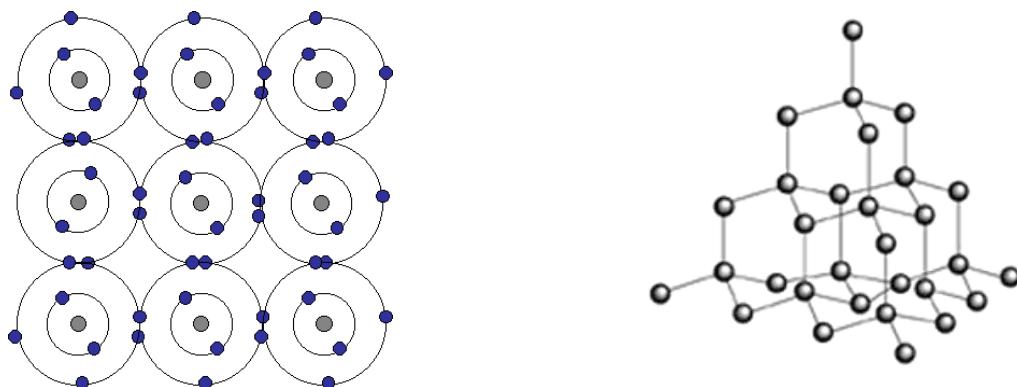


Figure (6.7) Covalent bonding in diamond

However, as the atoms come close, these energy levels start splitting to become a band. Thus 1s, 2s and 2p levels become 1s, 2s and 2p bands. As it can be noticed from the Fig (6.8), at first 2p level starts splitting and afterwards 2s and then 2p levels split. As the inter-atomic distance is further decreased, the electrons interact more strongly and thus the bands expand gradually. At a certain stage, 2s and 2p bands overlap (or merge), and thus a hybrid (composite) 2s-2p band is formed. The occupancy of this band is $2N + 6N = 8N$ states. However, it contains only $2N + 2N = 4N$ electrons. According to auf-bau rule, these electrons fill up the levels in the increasing order of energy. Thus the lower 4N states are filled at first and the upper 4N states remain empty. One may be tempted to conclude that, because of such partially filled band, diamond should behave as a conductor. However, the case is not so. As the inter-atomic distance is further decreased, the 2s-2p hybrid band further splits into two separate bands. Out of these, the lower 2s-2p occupied band possesses 4N states all filled with 4N electrons and the upper 2s-2p unoccupied band possesses 4N states all unoccupied. The lower 2s-2p band is thus the highest occupied and thus the valence band. The upper 2s-2p band is lowest unoccupied and thus the conduction band. The equilibrium inter-atomic distance of the diamond is 3.56 A° . When the atoms approach at this distance, the valence and conduction band continue to go away from each other. At 3.56 A° , the energy gap between these two bands becomes 5.7 eV . Thus at room temperature 300 K , the thermal energy of 0.038 eV is quite insufficient to excite the electrons in the conduction band. Thus diamond behaves as an insulator.

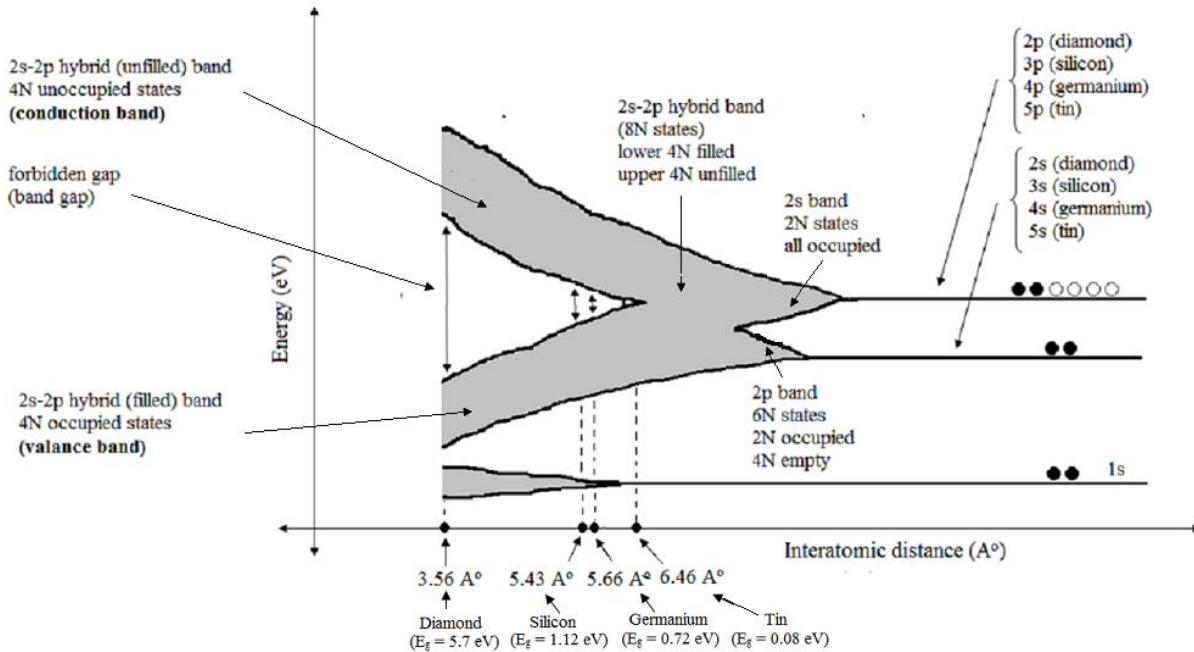


Figure (6.8): energy band formation in group IV elements

This 2s-2p hybridization and re-splitting effect also occurs in rest of the group IV elements such as silicon, germanium and tin. However, in silicon, germanium and tin, it occurs in 3s-3p, 4s-4p and 5s-5p levels respectively. We know that the energy levels are governed by $\frac{1}{n^2}$ rule. Thus the gap between the uppermost levels successively decreases as we move from diamond to silicon, germanium and tin. Further, as table (6.4) indicates, the inter-atomic distance increases as we approach from diamond to tin. Because of both these reasons, though all four elements have 2s-2p, 3s-3p, 4s-4p and 5s-5p hybridization, the gap between the valance and conduction band gradually decreases (refer Fig.6.8). Table (6.4) indicates that for silicon and germanium, these band gaps are 1.12 eV and 0.72 eV respectively, while for tin it is only 0.08 eV. Because of such extremely small band gap, tin behaves as a conductor. Silicon and germanium have intermediate band gaps and thus exhibit semiconducting properties. As it can be seen from the table (6.4), the atoms of these elements contain four valance electrons in the outermost orbit. Further, each atom is surrounded by four neighbors. An atom forms covalent bonds with the neighbors by sharing valance electrons to form covalent bonds (refer fig). Band theory works for crystals, which have periodic arrangement of atoms. The materials such as glass and plastics are insulators; however, band theory of solids is not applicable to them.

Insulators: Consider an example of diamond, which has a band gap of 5.7 eV. As we have seen earlier, the thermal energy $\left(\frac{3}{2}kT\right)$ of the valance electrons at 300 K is approximately 0.038 eV. This energy is thus lesser than that required to cross the band gap. This is quite insufficient for the valance electrons to cross the band gap and enter in the conduction band. It may be noted that the electron cannot be excited to any energy level within the band gap, as these energy levels are forbidden. Thus the concentration of the free electrons in diamond is extremely poor and consequently it is an insulator. Thus thermal energy at room temperature is not sufficient to

impart conducting properties to diamond. Can electrical energy be used to surmount the band gap then? One may be tempted to think that, for imparting an energy of 5.7 eV, potential difference of 5.7 volts (and the corresponding electric field) is sufficient. However, this imparted energy is immediately lost by the electron during its collisions with the crystal imperfections. It can be shown that an electric field of 10^8 V/m is required to impart 5.7 eV energy to the electron by considering the energy losses due to frequent collisions. This is another reason why diamond behaves like an insulator.

The typical band gap of diamond makes it transparent to the light. The band gap of diamond is 5.7 eV and the energies of the photons in the visible range is 1 to 3 eV. Thus light cannot be absorbed in diamond; therefore it is transparent to the light.

Diamond and graphite

We know that both diamond and graphite are the forms of carbon; however their electrical properties are quite different. Diamond is an insulator, while graphite is a good conductor of electricity. We know that carbon is tetravalent. In diamond, each carbon atom forms covalent bonds with its four neighbors. As all four electrons are utilized (localized) in strong covalent bonding, there are no free electrons in diamond and that's why it is insulator. In graphite, out of four valence electrons, only three valence electrons are localized in covalent bonding which occurs in a plane. The fourth electron is not involved in the covalent bonds and hence it is delocalized. Thus in graphite, every carbon atom has one delocalized electron which is free to conduct. Thus graphite is a good conductor of electricity.

Conductors: As we have discussed earlier, due to either partially filled valence band or overlapping of valence and conduction bands, there is no forbidden gap in the conductors. At absolute zero, the energy of electrons is not sufficient to enter the conduction band. As we shall see later, the highest occupied energy level of the electrons in the valence band is called as **Fermi level**. This level is the topmost energy level in the valence band. The electrons which are quite below the Fermi level cannot be directly excited to the conduction band, but at room temperature, due to thermal energy, the electrons just below the Fermi level can be easily excited to the conduction band. When electric field is applied, the electrons which are set free start moving and constitute the current. The resistivity of the conductors (metals) can be shown to be equal to

$$\rho = \frac{m}{ne^2\tau} \quad \dots(6.1)$$

Where, n is the number density of the electrons in the conduction band and τ is the relaxation time, that is the average time between the successive collisions. The electrons in the conductors also undergo frequent collisions with the lattice vibrations and as a result the relaxation time is considerably small ($\sim 10^{-14} \text{ s}$ in copper), however, the free electron density is extremely large ($\sim 10^{28} \text{ electrons/m}^3$ in copper), and thus the resistivity of conductors is extremely small.

In conductors, because of zero band gap, all valence electrons are set free at room temperature. When the temperature is raised further, this sea of electrons is thermally agitated.

This agitated sea of electrons offers resistance to the flow of conduction electrons. This is the reason why resistance of conductors increases with the temperature. Conductors thus have positive temperature coefficient of resistance.

When a conductor is exposed to light, it is absorbed because of zero band gap. Conductors are thus opaque to the visible light. The light sets the electrons free. Thus conductors can convert light into electricity. The best example of this phenomenon is photocell.

Semiconductors: The band gaps of silicon and germanium are 1.12 eV and 0.72 eV respectively. These are low as compared to the band gap of diamond but relatively higher than that of conductor. At room temperature, the thermal energy raises a small number of electrons in to conduction band. The free electron density in the silicon is roughly $10^{16} /m^3$, which is quite less as compared to copper ($10^{28} /m^3$). Due to these electrons, semiconductors exhibit a little-bit better conductivity than insulators. As we shall see later, the concentration of the charge carriers in the semiconductors can be increased significantly by adding impurities. Table provides list of semiconductors with their band gaps

It is possible to form compound semiconductors by using more than two elements. The elemental semiconductors are tetravalent. Compound semiconductors like GaAs, CdS etc. are formed by combining the elements from group III, V or group II, VI. The typical band gaps impart following useful properties to the semiconductors

Sr. No.	Name of the semiconductor	Symbol	Band gap (eV)	Wavelength corresponding to the band gap (A°)	Description
1	Silicon	Si	1.12	11099	Solar cells and other electronic components
2	Germanium	Ge	0.67	18553	Electronic devices
3	Silicon carbide	SiC	2.3	5404	Yellow LEDs
4	Gallium Nitride	GaN	3.44	3613	Electronics and spintronics, blue LEDs
5	Gallium Phosphide	GaP	2.26	5500	Red orange and green LEDs (after doping)
6	Gallium Arsenide	GaAs	1.43	8693	Next to silicon, IR LEDs, solar cells (after doping)
7	Cadmium sulphide	CdS	2.42	5136	Photoresistor, solar cells, quantum dots, lasers
8	Zinc Oxide	ZnO	3.37	3689	Laser diodes and conductive coatings

Table (6.4): List of semiconductors with their band gaps and applications

- i. **Semiconductors have two kinds of charge carriers:** When an electron is excited to the conduction band, it leaves a vacancy in the valance band. This vacancy is called hole. When a PD is applied, the hole moves opposite to electron and hence treated as a positive charge carrier. Thus semiconductors have bipolar conduction (electrons and

- holes). Holes can exist only in semiconductors and not in conductors and insulators. By doping, it is possible to make P and N type of semiconductors. Junctions made of P and N semiconductors are greatly used in electronics.
- ii. **Thermistors:** When heat is applied, the electrons are thermally excited to the conduction band. Thus resistance of the semiconductors decreases with temperature. Semiconductors are thus said to have negative temperature coefficient of resistance (NTC). Thermistors (temperature dependent resistors) find applications in measuring temperature, voltage regulation, circuit protection etc.
 - iii. **Semiconductors can convert light in to electricity and vice versa:** Table (6.5) shows that band gap of semiconductors are close to the photons of UV, visible and IR radiations. Thus semiconductors can absorb/emit light. Thus they are opaque to the light. This ability to convert light in to electricity (solar cells and photodiodes) and electricity in to light (LED and laser diodes) make them greatly applicable in optoelectronics (photonics)

Copper: The band gap of copper is 0 eV. Thus copper easily donates its valance electron in the conduction band. As copper is monovalent, each copper atom donates one electron in the conduction band. The density of the free electrons in copper is thus equal to its atomic density.

The atomic mass of copper is 63.

Thus 63 gm of copper contains 6.023×10^{23} atoms

Thus 1 gm of copper contains $\frac{6.023 \times 10^{23}}{63} = 9.56 \times 10^{21}$ atoms/gm

Density of copper is 8.96 gm/cm³

Thus 8.96 gm/cm³ of copper contains $9.56 \times 10^{21} \left(\frac{\text{atoms}}{\text{gm}} \right) \times 8.96 \left(\frac{\text{gm}}{\text{cm}^3} \right) = 8.57 \times 10^{22} \frac{\text{atoms}}{\text{cm}^3}$

Thus electron concentration of copper is $8.57 \times 10^{22} \frac{\text{electrons}}{\text{cm}^3} = 8.57 \times 10^{28} \frac{\text{electrons}}{\text{m}^3}$

6.5 FERMI DIRAC STATISTICS:

Statistical description of electrons in semiconductors

We know that when we consider a system of many particles (such as molecules, photons or electrons), we need a statistical approach to describe their properties. There are three such approaches, in all of which we use certain laws to describe a probability P(E) that a particular particle will have an energy E. There are three kinds of statistics that give such probabilities. These are

Maxwell-Boltzmann Statistics (Optional): This statistics is considered as a ‘classical’ statistics. It is mainly applicable to the molecules in the gas, which are considered as identical but distinguishable particles, whose wavefunctions do not overlap appreciably. As the distance between the particles is sufficient, the particles are distinguishable. The probability that an

average number of identical and distinguishable particles will have an energy E at temperature T is given by

$$P(E) = A e^{-\frac{E}{kT}} \quad \dots(6.2)$$

Where k is the Boltzmann constant and has value $1.3831 \times 10^{-23} \text{ J/K} = 8.617 \times 10^{-5} \text{ eV/K}$

In this statistics, there is no limit to the number of particles that can exist in a state of given energy. Maxwell Boltzmann statistics clearly explains raining which occurs due to evaporation of the water molecules on the sea surface. It also explains the sunshine, which occurs due to fusion of high energy protons in it. This statistics is applicable only in the situations where the temperature is high enough and density is low enough so that quantum effects can be ignored. As we shall notice it further, the other two quantum statistics, namely Fermi-Dirac and Bose-Einstein statistics approach Maxwell Boltzmann statistics when the temperature is high and density is low.

Bose Einstein statistics (Optional): This is a quantum statistics and is applicable to photons; the spin zero particles. It is also applicable to the particles having integer spin. Photons are also called bosons (in memory of Satyendranath Bose, an Indian Physicist) and they do not obey Pauli's exclusion principle. Photons are identical particles, but they are indistinguishable as their wavefunctions overlap considerably. The wavefunctions of photons do not change the sign if the bosons in any pair are exchanged and thus are symmetric. Any number of photons can occupy a quantum state of given energy. This statistics explain all properties of photons, for example the black body spectrum. The probability that a boson occupies a given quantum state of energy E is given by

$$P(E) = \frac{1}{A e^{\frac{E}{kT}} - 1} \quad \dots(6.3)$$

Fermi Dirac statistics: (Compulsory) This is also a quantum statistics which is applicable to all the particles having odd half integral spin ($\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \dots$). Electrons are the best example of the particles obeying this statistics. In this statistics, electrons are treated as waves, obeying Pauli's exclusion principle. Electrons obeying Pauli's exclusion principle are called fermions (in honor of Enrico Fermi). Electrons are identical but indistinguishable particles whose wavefunctions overlap considerably. The wavefunctions of electrons change the sign when electrons in any pair are exchanged and thus are called antisymmetric wavefunctions. A quantum state of given energy can be occupied only by only one fermion, which is the statement of Pauli's exclusion principle. The probability P(E) that a fermion occupies a quantum state of energy E, at the temperature T is given by

$$P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{kT}}} \quad \dots(6.4)$$



Enrico Fermi(1901-1954): He was one of the rare Physicists who excelled in both theoretical as well as experimental Physics. In his relatively short span of life of only 53 years, he made significant contributions in the fields of statistical mechanics, nuclear physics and elementary particle physics. He, along with Paul Dirac developed a statistics for electrons which obey Pauli's exclusion principle. This is known as Fermi Dirac statistics. Fermi is also known for his enormous contributions in nuclear physics. After the discovery of neutron by James Chadwick, Fermi attempted to produce transuranic elements by neutron bombardment and based on his results, he declared that slow neutrons were more effective in such reactions. This suggestion itself was responsible for discovery of highly explosive nuclear fission. Fermi became the first Physicist to build world's first nuclear fission reactor in 1942, just four years after discovery of fission. Fermi also contributed in the field of elementary particle physics and is known for the discovery of weak interactions, one of the four prominent interactions in the physical world. He also contributed in the experimental and theoretical physics of neutrino. He received Nobel prize in Physics 1938 for his significant contributions in nuclear physics. Indeed he is considered as father of nuclear engineering. Nobel prize is only one of the several awards and honors bestowed on him. Electrons which obey Fermi Dirac statistics are termed as fermions in his honor. A new transuranic element having atomic number 100 is named fermium in his honor. A characteristic energy level describing the behavior of electrons in metals and semiconductors is also called Fermi level in his honor. Fermi also received several patents in the field of nuclear physics and several institutions have been given his name, one being Fermi lab in USA.

Paul Dirac: (1902-1984): He was a theoretical physicist who made significant contributions in quantum mechanics and quantum electrodynamics. He was a professor in University of Cambridge. He, along with Fermi, formulated Fermi Dirac statistics which describes the behavior of electrons (called fermions: the spin $\frac{1}{2}$ particles) in metals and semiconductors. He formulated Schrodinger's equation with relativistic considerations which is now called Dirac equation. The solution of this equation predicted that electrons have spin and they have antiparticles. Later on both these predictions were experimentally confirmed. The antiparticle of electron is called positron and it was experimentally observed by Carl Anderson in 1932. He shared Nobel prize in Physics with Erwin Schrodinger in 1933. It is interesting to note that, the great Indian Physicist Homi Bhabha was his Ph.D. student. Although Paul Dirac obtained degree in electrical engineering, afterwards he turned to Physics during his doctoral studies. His Ph.D. thesis was based on quantum mechanics, the first ever thesis on this subject. He also proposed and investigated magnetic monopoles (yet to be observed). He also laid the foundations of superstring theory. His book named "Principles of Quantum Mechanics" is considered a landmark in the history of science. He also made significant contributions in quantum electrodynamics and also worked on quantization of gravitational field. He also provided theoretical basis for Pauli's exclusion principle.



Where k is the Boltzmann constant. E_F is called as Fermi energy. As we shall see in subsequent sections, Fermi energy plays a major role in determining the properties of metals, semiconductors and semiconducting devices such as diodes, transistors, solar cells etc. In any statistics, $P(E)$ can never exceed 1. If $E \gg kT$, then the Bose-Einstein and Fermi-Dirac distribution functions approach Maxwell Boltzmann distribution function. It is to be noted that the Fermi energy E_F appears only in Fermi Dirac function and not in others. The above formula clearly indicates that the probability of occupancy of higher energy levels increases with temperature. The

6.6 FERMI LEVEL:

A characteristic energy level which determines the behavior of semiconductors and semiconducting devices

The concept of Fermi level is essentially applicable to metals, semiconductors as well as insulators. Let us now try to find out the physical significance of the Fermi level. Let us consider eqn (6.4) once again. We have

$$P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{kT}}} \quad \dots(6.4)$$

Let us consider a situation where we have $T = 0$. In this situation, there are two possibilities; one $E > E_F$ and another $E < E_F$. Let us work out both

At $T = 0$ K and $E > E_F$ we have $(E - E_F) > 0$ (positive). Thus

$$P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{0}}} = \frac{1}{1 + e^{+\infty}} = 0$$

Thus at $T = 0$ K, the probability of occupancy of all the levels above the Fermi level is zero. Thus at $T = 0$ K, all the energy levels above the Fermi level are certainly unoccupied

At $T = 0$ K and $E < E_F$, we have $(E - E_F) < 0$ (negative). Thus

$$P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{0}}} = \frac{1}{1 + e^{-\infty}} = 1$$

Thus at $T = 0$ K, the probability of occupancy of all the energy levels below the Fermi level is one. Thus at $T = 0$ K, all the energy levels below the Fermi level are certainly occupied. Both these calculations indicate that at $T = 0$ K, Fermi level is the highest occupied energy level. All the states below Fermi level are occupied and those above Fermi level are unoccupied. This concept of the Fermi level is applicable to metals, semiconductors as well as insulators.

At $T = 0$ K and $E = E_F$ we have

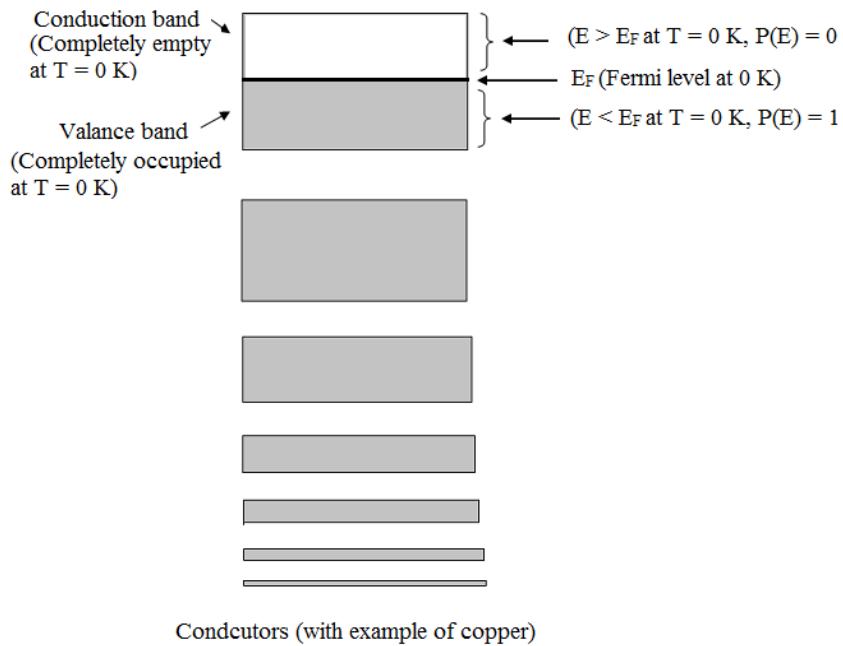
$$P(E) = \frac{1}{1 + e^{\frac{E - E_F}{kT}}} = \text{indeterminate}$$

Thus it is not possible to work out the situation, where we have $E = E_F$ and $T = 0$ K. Therefore let us consider a situation where we have $T > 0$ K and $E = E_F$. We get

$$P(E) = P(E_F) = \frac{1}{1 + e^{\frac{E_F - E_F}{kT}}} = \frac{1}{1 + e^0} = \frac{1}{1 + 1} = \frac{1}{2} = 0.5 = 50\%$$

Thus the probability of the occupancy of the Fermi level at any finite (nonzero) temperature is always 50%. Thus at any temperature higher than 0 K, Fermi level is the average energy level whose probability of occupancy is always 50 %, no matter what is the temperature (high or low). This also means that the physical significance of the Fermi level depends upon the temperature. At 0 K, it is the highest occupied level while at higher temperatures it is an average level having probability of occupancy 50 %, irrespective of value of the temperature. This is applicable to metals, semiconductors as well as insulators.

It is to be noted that when the temperature is raised, the electrons gain a kinetic energy given by kT . At $T = 300$ K, kT is 0.025 eV, while even at a very high temperature, say $T = 1000$ K (the temperature at which the metal glows brightly), kT is only 0.086 eV. This means that for temperatures greater than 0 K, (low or high) electrons gain an insignificant amount of energy to rise in the conduction band. The electrons which are at very low positions below Fermi level can thus not be excited to conduction band with such insignificant energy. These electrons can also not be excited to the energy levels just above them as those energy levels are already occupied by the other electrons. This means that at finite temperatures (low or high), only those electrons, which are very close below the Fermi level can be excited to the energy levels very close above the Fermi level. This discussion is summarized in Fig (6.9). Figures (6.10 and 6.11) show that when temperature is raised, the electrons just below the Fermi level vacate their levels and are excited to the energy levels just above the Fermi level. Thus the probability of the occupancy of the energy levels just below the Fermi level decreases and the probability of the occupancy of the levels just above the Fermi level increases. This corresponds to the curve corresponding to T_1 in Fig (6.11). It can be noted that now the curve intersects the Fermi level at $P(E) = 0.5$. When temperature is raised further, the process of decrease in the probability of occupancy of levels below the Fermi level and increase in the probability of occupancy of the levels above the Fermi level continues. However, at any temperature, low or high, the Fermi Dirac curve always passes through a point having coordinates $(E_F, 0.5)$. Thus at any finite temperature, Fermi level remains a level having probability of occupancy 0.5 (50%).



Conductors (with example of copper)

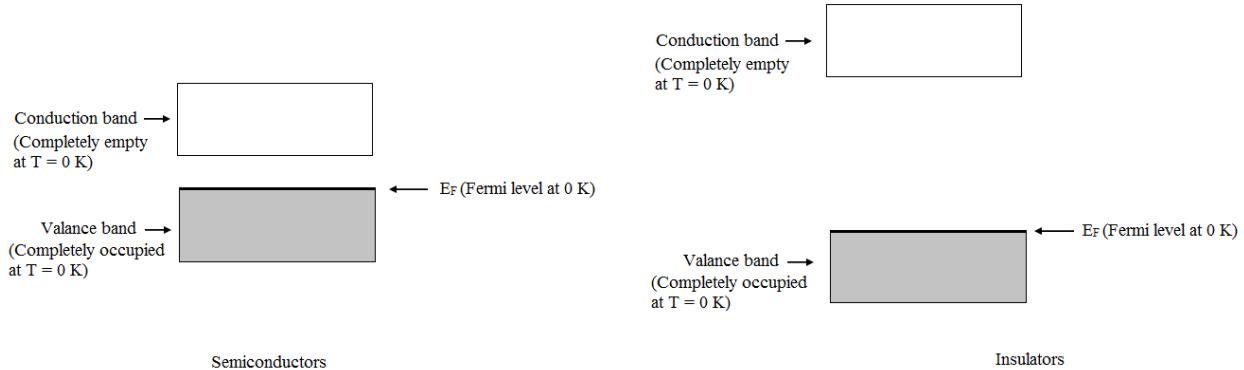


Figure (6.9): Physical significance of Fermi level at T = 0 K

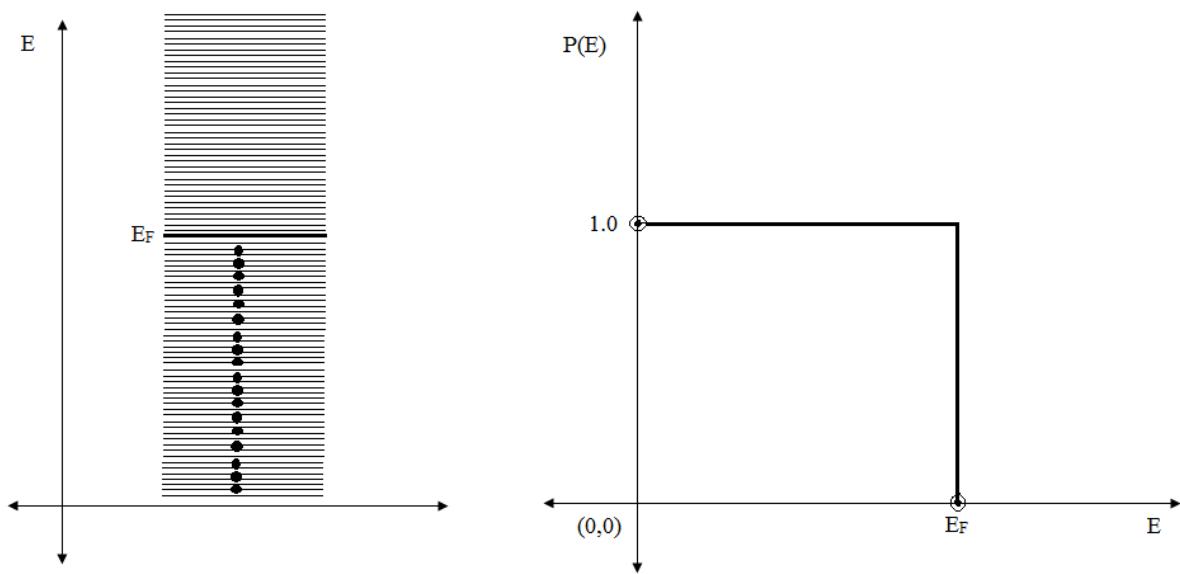


Fig (6.10): The significance of Fermi level in conductors (metals) at $T = 0 \text{ K}$

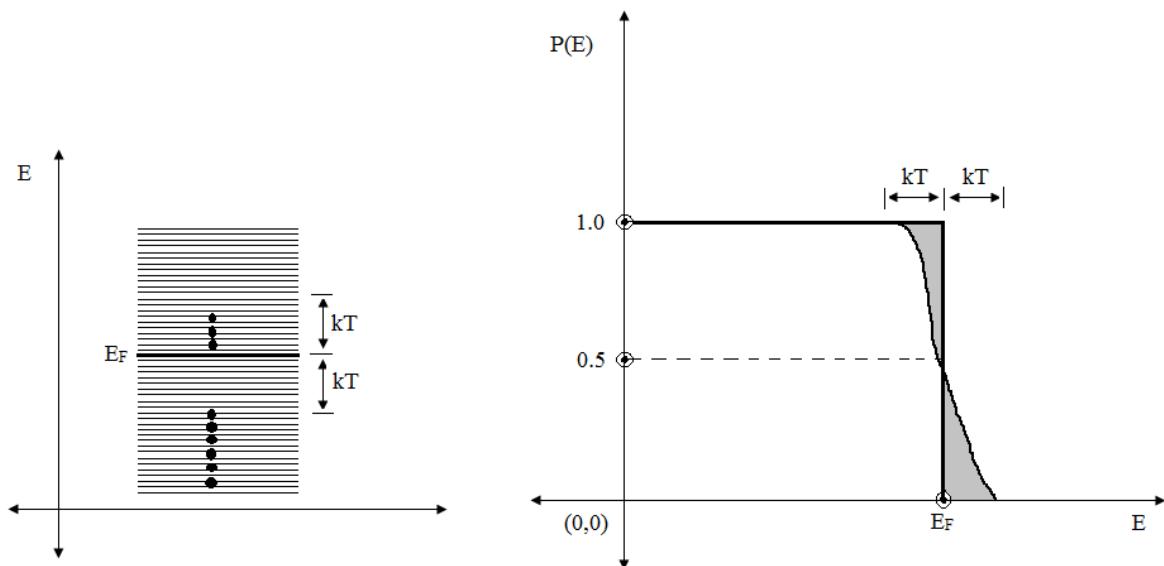


Fig (6.11): The significance of Fermi level in conductors (metals) at $T > 0 \text{ K}$

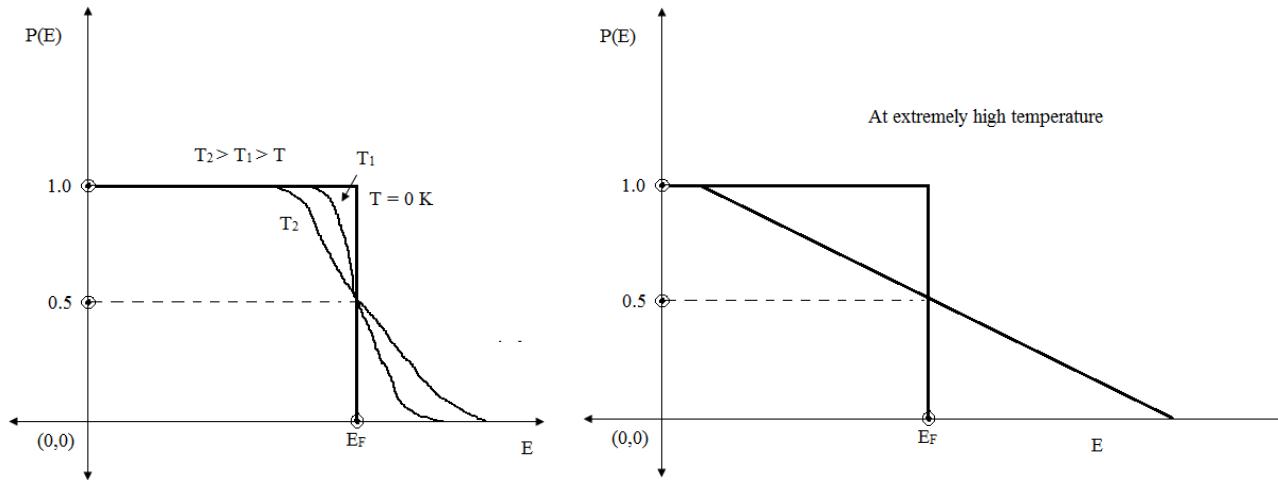


Fig (6.12): Graphical representation of Fermi Dirac statistics. Physical significance of the Fermi level

Fermi levels in conductors: further discussion (Optional)

The energy corresponding to Fermi level is called as Fermi energy and the speed corresponding to the Fermi energy is called Fermi speed. It can be shown that for copper the Fermi energy is 7.0 eV and the corresponding Fermi speed is $1.6 \times 10^6 \text{ m/s}$. Table 6.5) shows the Fermi energies in a few metals

Sr. No.	Metal	Symbol	Fermi energy (eV)
1	Lithium	Li	4.72
2	Sodium	Na	3.12
3	Aluminum	Al	11.8
4	Potassium	K	2.14
5	Cesium	Cs	1.53
6	Copper	Cu	7.04
7	Zinc	Zn	11.0
8	Silver	Ag	5.51
9	Gold	Au	5.54

Table (6.5): Fermi energies in some metals

It is possible to calculate the number of quantum states $n(E)$, per unit energy interval. This called as density of states and is given by

$$n(E) = \frac{8\sqrt{2}\pi m^{\frac{3}{2}}}{h^3} E^{\frac{1}{2}} \quad \dots(6.5)$$

A plot of $n(E)$ Vs E is shown in Fig (6.13 a, b, c). Note that the Fig (a) indicates density of states $\{n(E)\}$ while Fig (c) indicates density of occupied states $\{n_o(E)\}$. $n_o(E)$ is the product of $n(E)$ and $P(E)$. Thus the graph in Fig (c) is the product of graphs in Fig (a) and Fig (b)

Using Eqn (6.5) and the graphs, it is possible to formulate the following equation enabling us to calculate the Fermi energy.

$$E_F = \frac{0.121h^2}{m} n^{\frac{2}{3}} \quad \dots(6.6)$$

Where n is the number of electrons per unit volume. The Fermi energies calculated for various metals is given in table (6.5)

Fig (6.14 a, b and c) depict the picture at $T > 0$. The probability of occupancy of the levels above the Fermi level increases at the cost of probability of energy levels below E_F . Thus the density of the occupied states above E_F increases at the cost of density of occupied states below E_F . Note that the pictures at $T > 0$ are very close to pictures at $T = 0$, indicating that even at higher temperatures, modifications take place only in the close vicinity of the Fermi level.

Example (6.1) (Optional): Calculate the speed of conduction electron in copper having its kinetic energy equal to Fermi energy of 7.0 eV. Also calculate the drift velocity if a current of 5 A is flowing in a copper wire with a cross section of 0.5 mm^2 . Free electron density in copper is $8.5 \times 10^{28}/\text{m}^3$ and the charge on the electron is $1.6 \times 10^{-19} \text{ C}$. Which velocity is greater? Why? Comment on the results

Solution:

$$\text{We have } KE = \frac{1}{2}mv^2$$

$$\Rightarrow 7 \text{ eV} = \frac{1}{2}mv^2$$

$$\Rightarrow 7 \times 1.6 \times 10^{-19} J = \frac{1}{2} \times 9.1 \times 10^{-31} v^2$$

$$\Rightarrow v = 1.6 \times 10^6 \text{ m/s}$$

$$\text{Now, } I = nev_d A$$

$$\Rightarrow 5 = 8.5 \times 10^{28} \times 1.6 \times 10^{-19} \times v_d \times 0.5 \times 10^{-6}$$

$$\Rightarrow v_d = 7.4 \times 10^{-4} \text{ m/s}$$

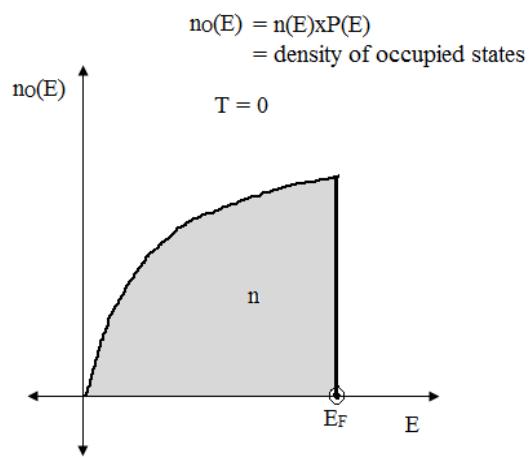
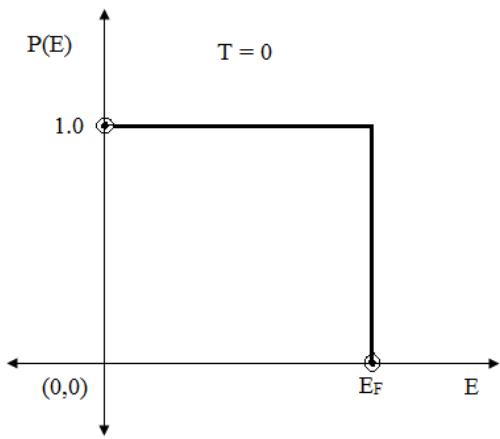
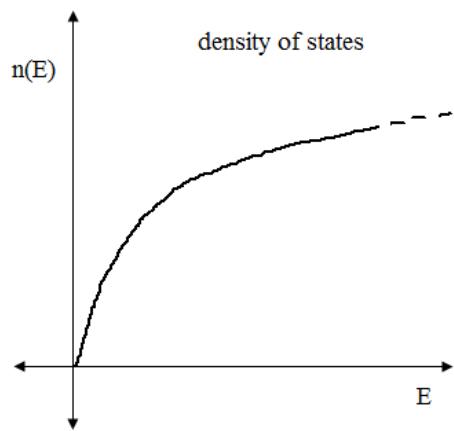


Figure (6.13 a, b, c): Density of states at $T = 0$ K

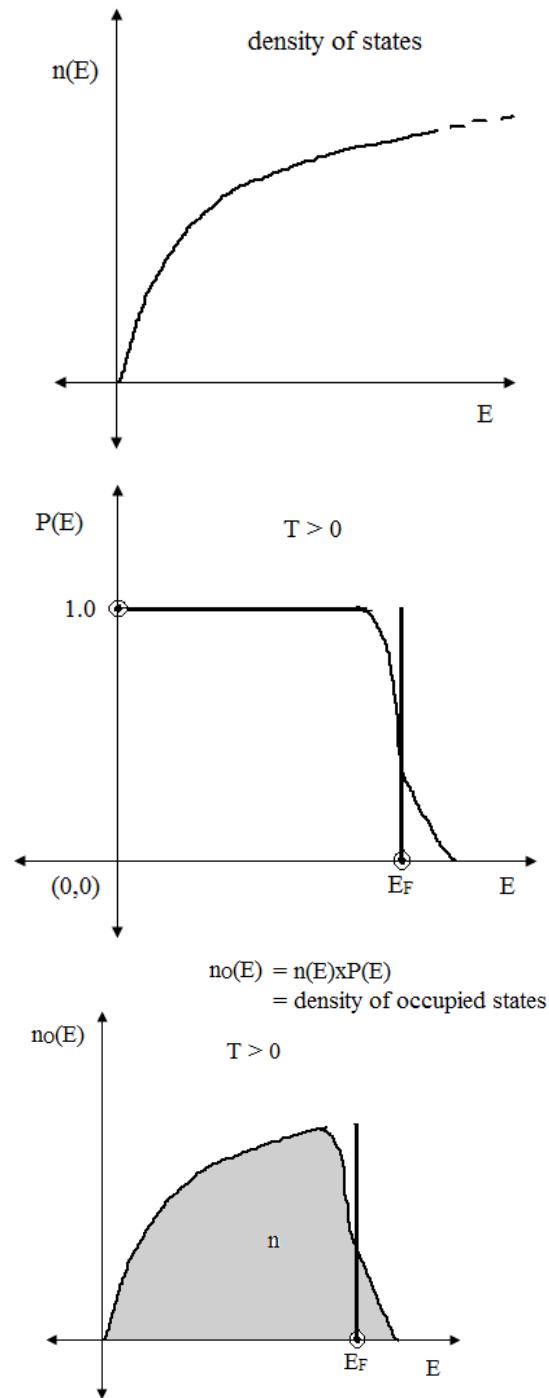


Figure (6.14): Density of states at $T > 0$ K

Fermi level in intrinsic semiconductors at T > 0: Further discussion (Compulsory):

The drift velocity of electrons is considerably less as compared to instantaneous velocity (velocity acquired between the successive collisions). This is because drift motion is considerably affected due to the collisions. The semiconductors in purest form (without any dopant) are called intrinsic semiconductors. We know that semiconductors have a moderate band gap ($E_g \approx 1 \text{ to } 3 \text{ eV}$). We have also seen that at 0 K, the Fermi level is the highest occupied level and it is situated on the top of the valence band. We have also seen that at any finite nonzero temperature, the Fermi level is an average energy level whose probability of occupancy is always 50%, irrespective of the value of the temperature. This is applicable to conductors, semiconductors as well as insulators. Thus in intrinsic semiconductors also, the Fermi level at finite temperature is an average energy level, whose probability of occupancy is always 50%. A relation consistent to this discussion can be written as

$$E_F = \frac{E_V + E_C}{2} \quad \dots(6.7)$$

Where E_V and E_C are the average energies of the electrons in valance and conduction bands respectively.

Where such level can be shown in the energy band diagram then? Commonsense as well as the above relation tells us that it should be situated in the middle of the forbidden band. Eqn. (6.7) can be proven mathematically

(This derivation is optional)



According to Fermi Dirac statistics, the probability $P(E)$ of occupancy of an energy level E is given by

$$P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{kT}}} \quad \dots(6.8)$$

Thus the probabilities of occupancies of the energy level in valance band and conduction band are

$P(E_V) = \frac{1}{1 + e^{\frac{(E_V-E_F)}{kT}}}$	and	$P(E_C) = \frac{1}{1 + e^{\frac{(E_C-E_F)}{kT}}}$... (6.9)
---	-----	---	-----------

The electron can either be found in the valance band or the conduction band. Further, the total probability of finding the electron in valance and conduction band is always 1. Thus we have

$$P(E_V) + P(E_C) = 1$$

Substituting $P(E_V)$ and $P(E_C)$

$$\begin{aligned}
& \frac{1}{1 + e^{\frac{(E_V - E_F)}{kT}}} + \frac{1}{1 + e^{\frac{(E_C - E_F)}{kT}}} = 1 \\
& \Rightarrow \frac{1 + e^{\frac{(E_C - E_F)}{kT}} + 1 + e^{\frac{(E_V - E_F)}{kT}}}{\left(1 + e^{\frac{(E_V - E_F)}{kT}}\right) \times \left(1 + e^{\frac{(E_C - E_F)}{kT}}\right)} = 1 \\
& \Rightarrow 1 + e^{\frac{(E_C - E_F)}{kT}} + 1 + e^{\frac{(E_V - E_F)}{kT}} = \left(1 + e^{\frac{(E_V - E_F)}{kT}}\right) \times \left(1 + e^{\frac{(E_C - E_F)}{kT}}\right) \\
& \Rightarrow 1 + e^{\frac{(E_C - E_F)}{kT}} + 1 + e^{\frac{(E_V - E_F)}{kT}} = 1 + e^{\frac{(E_C - E_F)}{kT}} + e^{\frac{(E_V - E_F)}{kT}} + \left(e^{\frac{(E_V - E_F)}{kT}}\right) \times \left(e^{\frac{(E_C - E_F)}{kT}}\right)
\end{aligned}$$

Cancelling the common terms

$$\Rightarrow 1 = e^{\frac{E_V + E_C - 2E_F}{kT}}$$

Taking logarithms on both sides

$$\begin{aligned}
& \Rightarrow 0 = \frac{E_V + E_C - 2E_F}{kT} \\
& \Rightarrow E_F = \frac{E_V + E_C}{2} \quad \dots(6.10)
\end{aligned}$$

This clearly indicates that in intrinsic semiconductors, the Fermi level is situated in the middle of the band gap (refer Fig 6.5).

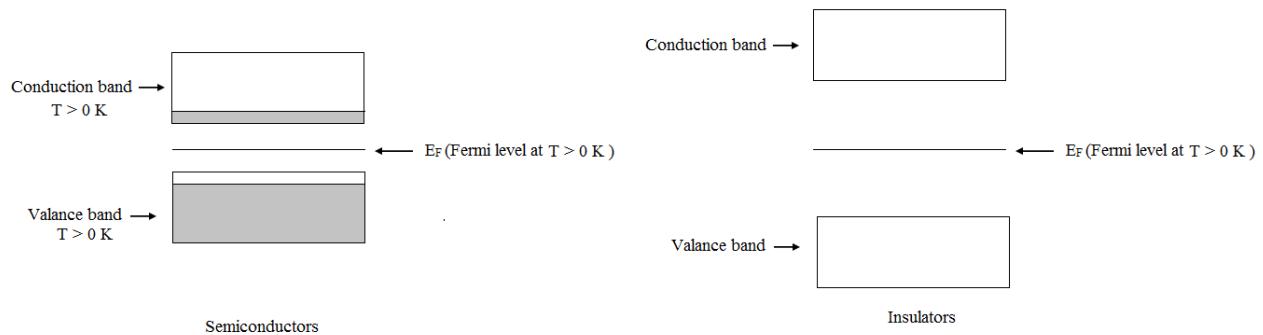


Figure (6.5): Fermi level in intrinsic semiconductors and insulators at $T > 0$ K

Thus we have another ‘definition’ of Fermi level. Fermi level in semiconductors is a level which corresponds to center of gravity of electrons in conduction band and holes in valance band when

weighted according to their energies. Note that all levels in the band gap are the forbidden levels. Thus fermi level is also a forbidden level. How a level having probability of occupancy 50% can be a forbidden level? This can be understood with two analogies. We know that the center of gravity of a hollow sphere is situated at its center, and there is no mass at the center. It is possible that the average result of a class of students is 50% but there is no student having scored exactly 50%. We can also extend an analogy of results to interpret the two definitions of Fermi level. Consider result of class in which the highest score is 77%. It is possible to gauge the quality of the students just by looking the highest score. For example if the highest score of the class is 57% in one case and 77 % in another, then the later one can be considered a class with better quality. This is analogous to definition of the Fermi level at $T = 0$ K. Now suppose, the teacher is not happy with the result of the class with highest score 57%, and thus he works hard on the class, and ‘excites’ many students to the higher scores. This can lead to a situation, where, now 57% is not the topmost score but an average score of the class. Being an average score, it’s probability 50%, but it is possible that no student has scored exactly 50%. This is analogous to the interpretation of the Fermi level at higher temperatures.

6.7 INTRINSIC SEMICONDUCTORS:

Semiconductors in purest form, but without applications

We have seen that semiconductors behave as insulators at 0 K. This because electrons in the valance band have no energy to cross the band gap. At room temperature a small number of electrons pick up thermal energy and are excited to the conduction band. The conductivity of semiconductors at room temperature is better than that of insulators, but inferior to the conductors

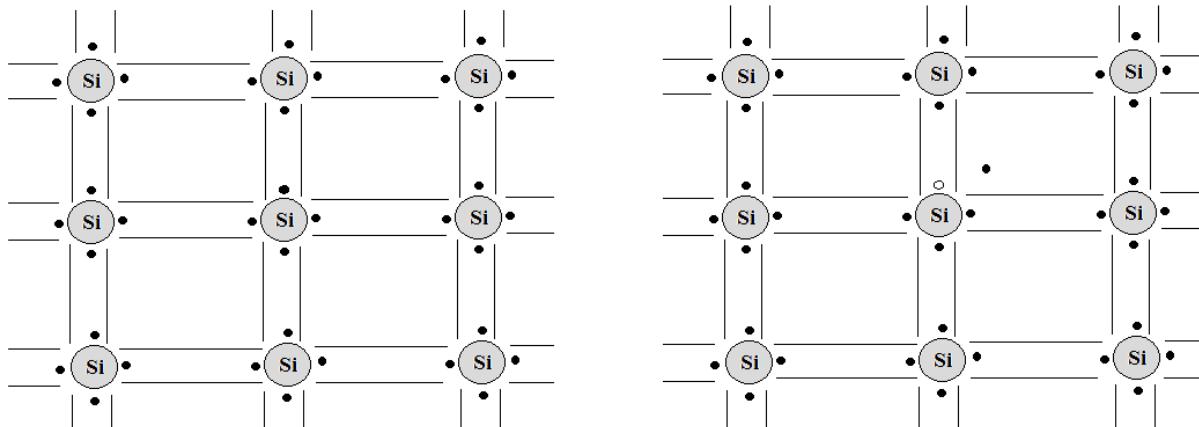


Figure (6.6): Silicon crystal at $T = 0$ K and $T > 0$ K

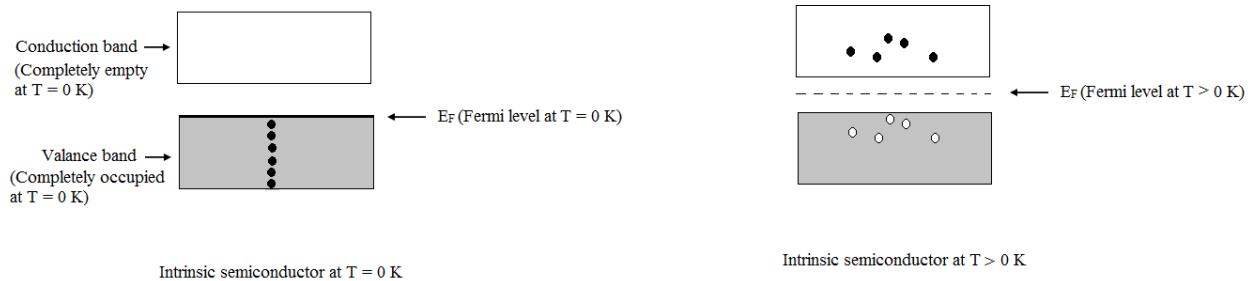


Figure (6.7): Energy band in silicon at $T = 0\text{ K}$ and $T > 0\text{ K}$

There are two prominent semiconductors, namely silicon ($Z = 14$, $E_g = 1.12\text{ eV}$) and germanium ($Z = 32$, $E_g = 0.72\text{ eV}$). They have a tetrahedral crystalline structure much similar to that of diamond. Consider silicon. Each atom of silicon is surrounded by four silicon atoms (Fig 6.6). As seen earlier, each silicon atom has four valence electrons. The valence electrons of neighboring silicon atoms are shared in covalent bond. As covalent bonds are stable, these electrons exist in valence band. At 0 K , all the electrons are in valence band. When temperature is raised, a small number of electrons gain sufficient energy to rise in conduction band. This corresponds to breaking of a covalent bond. When an electron is detached from the covalent bond, a vacancy is created. This vacancy can be conveniently treated as a hole. The vacancy makes the covalent bond unstable. This instability creates a location where one electron is greatly needed. Due to this and since the vacancy of electron corresponds to a positive charge; the vacancy is equivalent to a place with positive charge. The covalent bond with such vacancy can become stable only if another electron fills the vacancy. Two kinds of electrons can fill up the vacancy. One is free electron in conduction band. Such electrons move randomly in the free space in the crystal. If such electron suffers a collision with the atom, it loses the energy and thus falls in the valence band. Then it fills the vacancy. If this happens then the electron-hole pair disappears. This also results in the emission of a photon having energy, $h\nu = E_c - E_v$. Another possibility is that a bound electron in the neighboring covalent bond, say from site B may jump into the vacancy (hole) at the site A. If this happens then the hole at the original site A disappears, but a vacancy is created at site B. thus when a valence electron moves from site B to A, the hole moves from site A to B. Thus the movement of the hole is opposite to that of electron. It is in this sense that hole is considered as a positive charge carrier. When electric field is applied, the covalently bonded valence electrons move from one hole to another and thus hole moves from one site to another. When electric field is applied, these holes move in a direction exactly opposite to that of the electrons in the conduction band. Thus holes move in the direction of the electric field. Thus, under the application of the electric field, the semiconductor conducts due to two kinds of electrons, one, the free electrons in the conduction band (opposite to applied electric field) and another, the covalently bonded electrons in the valence band, the flow of which can be conveniently represented by the holes (along the electric field). It is to be noted that holes move through the valence band while the free electrons move through the conduction band. Holes cannot exist in conduction band. The motions of electrons and holes are opposite to each other. Thus a characteristic feature of the semiconductors is that they conduct due to two kinds of charge carriers; electrons and holes. As we shall see later, it is possible to enhance the

concentration of electrons as well as holes by doping. Thus there are two types of semiconductors, N type (mainly conducts due to electrons) and P type (mainly conducts due to holes). A variety of electronic devices such as PN junction diodes, NPN or PNP transistors, solar cells, photodiodes, Zener diodes can be made by using an appropriate combination of N and P types of semiconductors. Note that this is not possible by using any conductor, though it has better conductivity than semiconductor. Conductors conduct due to only one kind of charge carriers; electrons.

Now let us discuss about the concentration of charge carriers in the semiconductors. This can be calculated by using following method. Number of atoms can be calculated by

$$N = \frac{N_A \rho}{M} \quad \dots(6.11)$$

Where N_A is Avogadro's number, ρ is density and M is the atomic weight. Thus for silicon, we have

$$n = \frac{N_A \rho}{M} = \frac{\left(6.023 \times 10^{26} \frac{\text{atoms}}{\text{k.mole}}\right) \times \left(2330 \frac{\text{kg}}{\text{m}^3}\right)}{\left(28.09 \frac{\text{kg}}{\text{k.mole}}\right)} = 5 \times 10^{28} \frac{\text{atoms}}{\text{m}^3}$$

Not all, but only a small number of electrons from these many atoms can be excited into conduction band. Calculations show that this number is 10^{16} electrons/m³. These are the free electrons in the conduction band. As each electron raised to conduction band leaves a hole in the valence band, the intrinsic semiconductor contains an equal number of holes in the valence band. Note that, all these charge carriers are generated due to thermal energy. The word 'intrinsic' indicates that such semiconductors conduct due to their own thermally generated charge carriers. As we shall see later, such thermally generated carriers become minority charge carriers in extrinsic semiconductors.

As seen earlier, the Fermi level in intrinsic semiconductors at any finite nonzero temperature is at the center of the forbidden gap. This indicates that for any finite temperature, the concentrations of free electrons in conduction band and holes in valence band are always equal. If effective masses of electrons and holes are considered, then the Fermi level shows a minor shift in the original position with the rise in temperature. However, this shift is insignificant.

Limitations of intrinsic semiconductors:

Intrinsic semiconductors cannot be used directly in the electronic circuits as their conductivities are very low (10 million times weaker as compared to conductors). Secondly the conductivity strongly depends upon the temperature (in an exponential manner), indicating that a minor rise in the temperature will alter the electrical properties considerably. The conductivity cannot be modified by external means and the current depends more significantly on temperature than the voltage

6.8 EXTRINSIC SEMICONDUCTORS:

How to enhance the conductivity of semiconductors

For the semiconductors to be useful in electronics, one needs to enhance their conductivity and make them stable as regards to the temperature. This can be done very effectively by adding a minute and controlled amount of the impurity in the lattice of the intrinsic semiconductor. The process of adding the impurity is called doping. As we shall see later, this drastically improves the electrical conductivity of the semiconductor. The percentage of addition of the impurities is very small. For ex. there is only one impurity atom in 100 billion atoms of the intrinsic semiconductor. (10^8 host atoms: 1 impurity atom). As we shall see later, even such minute impurity can enhance the conductivity of the semiconductor by a factor of 10000! Such semiconductors are called as extrinsic semiconductors and they are superior to the intrinsic semiconductors as regards to better conductivity and it's stability against the temperature variations. Such extrinsic semiconductors are enormously applicable in electronics. The reason for keeping the dopant concentration very low is that the electrical conductivity is to be enhanced by keeping the crystal structure of the host almost intact. One of the methods of doping is to add the impurities in the host in its molten state. When the host crystal is grown, some of the host atoms are replaced by the impurity atoms. Two types of impurities, namely pentavalent and trivalent are added and this results in to two types of extrinsic semiconductors namely N and P types of semiconductors. The size of impurity atoms is almost same as of atoms in host and thus they 'fit' perfectly in to the lattice

N type Semiconductor: Table (6.6) shows the pentavalent impurities and their electronic configurations. It can be observed that the outermost orbit of all these elements contains 5 electrons. Thus, if phosphorous is doped in silicon, it will replace some of the silicon atoms from their site. As the doping is weak, there is strong possibility that almost all phosphorous atoms will be surrounded by 4 silicon atoms. Thus the 4 out of 5 electrons of phosphorous are shared in

Element	Electronic configuration
Phosphorous (P, Z = 15)	$1s^2 2s^2 2p^6 \mathbf{3s}^2 \mathbf{3p}^3$
Arsenic (As, Z = 33)	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} \mathbf{4s}^2 \mathbf{4p}^3$
Antimony (Sb, Z = 51)	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} \mathbf{5s}^2 \mathbf{5p}^3$
Bismuth (Bi, Z = 83)	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^6 5d^{10} 5f^{14} \mathbf{6s}^2 \mathbf{6p}^3$

Table (6.6): The electronic configurations of the pentavalent elements

covalent bonds, but one electron remains unshared. As covalent bonds are strong, the corresponding electrons are tightly bound and they thus exist in valance band of the silicon. However, the unshared electron is almost free. Note that if phosphorous atoms were in open space, even this fifth electron would have been as tightly bound as remaining four electrons. However the phosphorous atom in the host is 'alone' and cannot hold this fifth unshared electron tightly. It requires very less amount of energy to detach it (0.05 eV in silicon and 0.01 eV in

germanium). At room temperature thermal energy is quite sufficient to raise this electron in to the conduction band. As every phosphorous atom ‘donates’ one free electron, it is called as

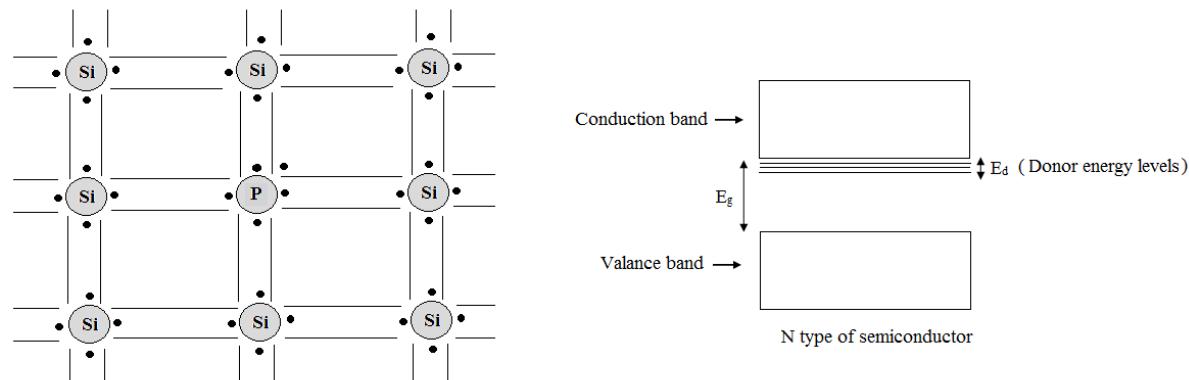


Figure (6.8): N type of semiconductor and it's energy band diagram

‘donor impurity’. Unlike the 4 shared electrons, the energy levels of unshared electrons cannot be represented in the valance band, as it would mean that their energy gap is 1.1 eV. The negligible energy required to detach the unshared electrons indicates that the energy levels of these electrons be shown just below the conduction band. These energy levels are called donor levels (Fig 6.8). They are discrete because the negligible interactions of the donor atoms which are quite distant from each other. At room temperature, the thermal energy of approximately 0.025 eV is quite sufficient to excite all the electrons in the donor levels in to the conduction band. Note that at room temperature itself, a small number of electrons in the valance electrons are also excited to the conduction band. Such electrons leave holes behind them. The electrons excited from donor levels do not leave any hole behind them. As we shall calculate in a sample problem, the concentration of the silicon’s own electrons in conduction band (and holes in the valance band) prior to doping is $10^{16}/\text{m}^3$. However after doping, additional electrons from donor energy levels are added in to the conduction band. As we shall prove it later, the concentration of these donated electrons is $10^{22}/\text{m}^3$. Thus $10^{22}/\text{m}^3$ donated electrons (without any equivalent holes) are added to $10^{16}/\text{m}^3$ preexisting electrons (with 10^{16} equivalent holes in valance band). The sum of 10^{22} and 10^{16} is almost 10^{22} . This means that after doping, there are $10^{22}/\text{m}^3$ electrons in conduction band and 10^{16} holes in valance band. Thus for every hole in valance band there are one million electrons in conduction band. Therefore in such cases electrons are called majority charge carriers while holes are called minority charge carriers. As the conduction in such semiconductors is mainly due to electrons, they are called N type of semiconductors. Note that the word ‘extrinsic’ indicate that the conduction is mainly due to the externally added charge carriers. We have also seen that Fermi level is a reference level which is equivalent to center of gravity of electrons in the conduction band and holes in valance band (when weighed with respect to energies), it is obvious that the Fermi level of N type of semiconductors is close to the conduction band.

P type semiconductors: The doping process can also be used to enhance the concentration of holes in the semiconductors. Let us consider trivalent impurities and their electronic

configuration as shown in table (6.7). As it can be noted from the table, all these elements contain 3 valence electrons in the outermost orbit. Consider that aluminum is doped in silicon.

Element	Electronic configuration
Aluminum (Al, Z = 13)	$1s^2 2s^2 2p^6 3s^2 3p^1$
Gallium (Ga, Z = 31)	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^1$
Indium (In, Z = 49)	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^1$
Thallium (Tl, Z = 81)	$1s^2 2s^2 2p^6 3s^2 3p^6 3d^{10} 4s^2 4p^6 4d^{10} 4f^{14} 5s^2 5p^6 5d^{10} 5f^{14} 6s^2 6p^1$

Table (6.7): the electronic configuration of the trivalent impurities

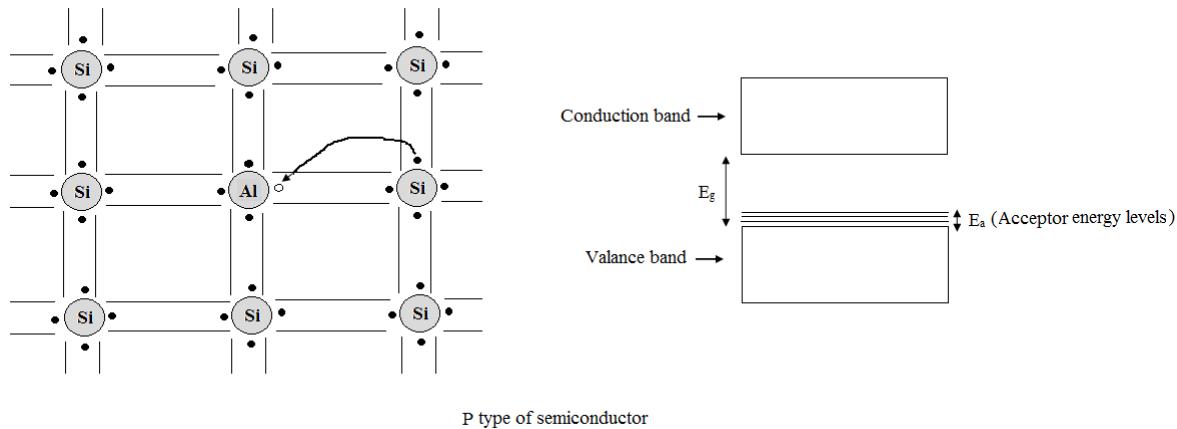


Fig (6.9): P type semiconductor and its energy band diagram

Now an aluminum atom is surrounded by 4 silicon atoms. Thus during the establishment of the covalent bonds, 3 electrons of aluminum are shared in the covalent bonds, while the fourth covalent bond is devoid of one electron. As we have seen earlier, such vacancy in the covalent bond corresponds to hole and it behaves like a positively charge carrier. As covalent bonds with vacancy are unstable, there is a great need of the electron there. The hole is thus ready to ‘accept’ the electron. Therefore the trivalent impurities are called as ‘acceptor’ impurities. The hole can be filled by the covalently bonded electrons in the neighboring sites. These electrons are situated in valance band. When a covalently bonded electron occupies a hole, its energy is slightly increased. Thus due to addition of acceptor impurities, additional energy levels are created just above the valance band. These levels are called as acceptor levels and they are occupied by the holes created due to doping process. The electrons in the valance band require very little energy to fill these holes. This energy is easily available at room temperature. When the electrons in the valance band are accepted in the acceptor levels, they leave holes in the valance band behind them. Thus acceptor energy levels are those which are supposed to be occupied by the electrons in the valance band which fill up the holes (Fig 6.9). As we have seen earlier, there are originally $10^{16}/m^3$ thermally excited electrons (and equivalently $10^{16}/m^3$ holes in the valance band) in the host. Due to doping, the number of holes in the valance band increases to $10^{22}/m^3$. Note that for such holes, there are no equivalent electrons in the conduction band. Thus the situation discussed previously occurs here also. There are 10^{16} preexisting thermally excited electrons in

the conduction band and 10^{22} ($10^{16} + 10^{22} \approx 10^{22}$) holes in the valance band. Thus for every electron in the conduction band, there roughly one million holes in the valance band. Thus in this case holes are majority charge carriers and electrons are minority charge carriers. As the conduction in such semiconductors is mainly due to the holes, they are called as P type of semiconductors. The Fermi level ('center of gravity') thus shifts close to the valance band.

6.8 DESCRIBING SEMICONDUCTORS BY MEANS OF FERMI LEVEL

Fermi level of a semiconductor is governed by its type, temperature and doping level

We have seen that in intrinsic semiconductors Fermi level lays in the center of band gap indicating equal concentration of electrons in conduction band and holes in valance band. However, the picture is different for extrinsic semiconductors. As the Fermi level is analogous to center of gravity of electrons in conduction band and holes in valance band (weighed with respect to energy), we expect the Fermi level to be close to conduction band in N type semiconductors and close to valance band in P type semiconductors. However this is an incomplete argument and we need to know, 'how much close? In fact the 'closeness' of Fermi level to either band is governed by temperature and doping level.

Fermi level in N type semiconductors:

We know that in N type semiconductor; there are donor energy levels just below the conduction band. At 0 K, these levels are occupied by unshared electrons of pentavalent impurity. Majority of the electrons in the conduction band come from these donor levels. At 0 K, all these electrons exist in donor levels only. Therefore the Fermi level of N type of semiconductor at 0 K is situated in the middle of the donor levels (refer Fig 6.10). Thus, we have

$$E_C \geq E_{FN\ T=0K} \geq E_D$$

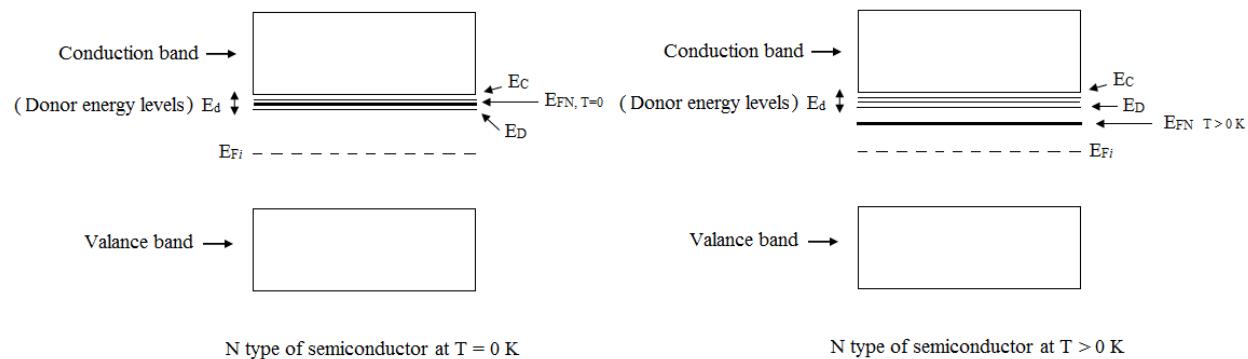


Figure (6.10): Fermi level in N type semiconductor at $T = 0 K$ and $T > 0 K$ (moderate)

Where E_C is the lowest energy level in the conduction band and E_D is the lowest energy level amongst the donor levels. This can also be expressed as

$$E_{FN\ T=0K} = \frac{E_C + E_D}{2}$$

When temperature is raised, the electrons in the donor levels start gaining thermal energy and they move to conduction band. Thus the donor levels are gradually depleted. As a result, the Fermi level starts moving down. As temperature is further raised, the electrons in the valance band also start going in to the conduction band. This results in generation of holes in valance band. To account for this, the Fermi level further shifts down. However at moderate temperatures, the Fermi level can never reach the intrinsic Fermi level, as it would wrongly indicate the equal concentration of the electrons and holes. Thus at moderate temperatures, Fermi level of N type of semiconductors (E_{FN}) is somewhere in between the lowest donor level and intrinsic Fermi level (Fig 6.10). Thus we have

$$E_D \geq E_{FN\ T > 0K} \geq E_{Fi}$$

As temperature reaches extremely high values, almost all electrons in the valance band are excited to the conduction band. These electrons are extremely large in number as compared to the electron donated by the donor levels. Thus at very high temperatures, the concentration of the electrons in conduction band and holes in the valance band is almost equal. This indicates that at very high temperatures, N type semiconductor is converted to intrinsic semiconductor. In this situation the Fermi level of the N type of semiconductor coincides with the intrinsic Fermi level. The gradual variation of the Fermi level with temperature is shown in Fig ...

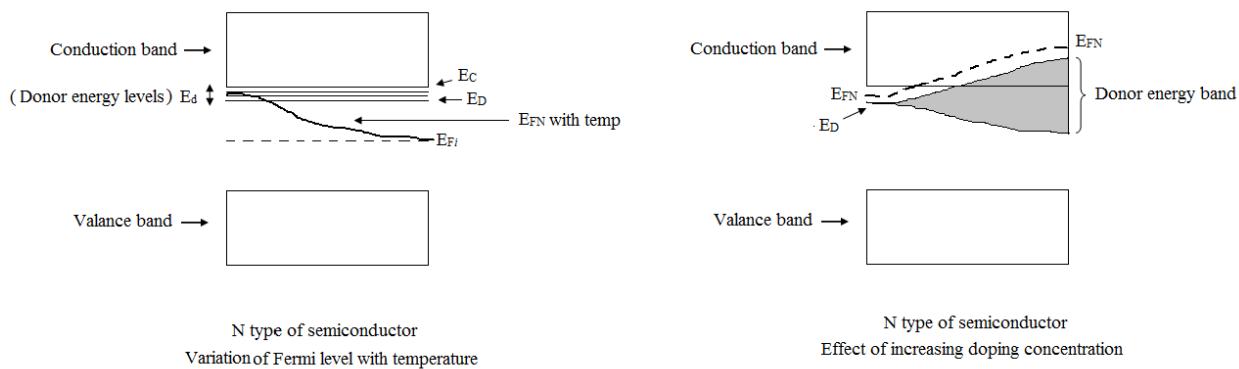


Figure (6.11) : Effect of temperature and dopant concentration on Fermi level of N type semiconductor

What happens if the concentration of the dopant (pentavalent) is increased? Note that if the concentration is moderate, then the distance between the donor atoms is large enough. As a result the interaction between the donor atoms is also weak. Therefore the donor levels do not split and they remain discrete. However, as the concentration is further increased the distance between the donor atoms decreases and they start interacting. Consequently, the donor levels interact and they become a donor band. If the concentration is still increased, donor levels split to a greater

extent and the donor band expands and invades in the conduction band. As a result the Fermi level also enters in the conduction band (Fig 6.11)

Fermi level in P type semiconductors:

The discussion of Fermi level in P type of semiconductor is opposite to N type semiconductor. Refer Fig (6.12)...

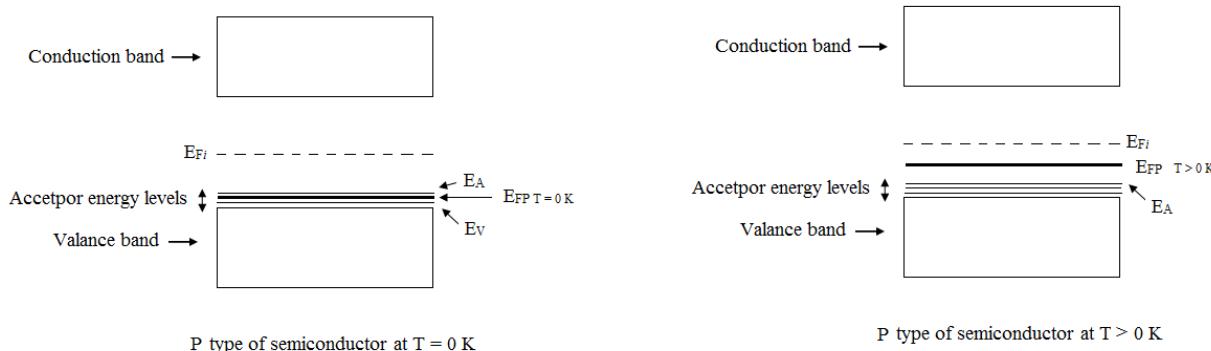


Figure (6.12): Fermi level in P type of semiconductors

We know that in P type semiconductor; there are acceptor levels near the valance band. The holes in P type semiconductor are only due to transitions of electrons from valance band to the acceptor levels. As the valance band is the source of electrons and acceptor levels are recipients of them, the Fermi level is situated in the middle of the group of acceptor levels (Fig 6.12). Thus we have

$$E_A \geq E_{FP\ T=0K} \geq E_V$$

Where E_A is the highest acceptor level and E_V is the highest level in the valance band. When temperature is raised, the electrons in the valance band gain thermal energy to raise in the acceptor levels. As a result, concentration of electrons in the acceptor levels increases and the Fermi level moves upward. As the temperature is further increased, at moderate temperatures, the electrons in the valance band are thermally excited to the conduction band. Thus, concentration of the electrons in the conduction band increases. As a result, the Fermi level further shifts up (Fig 6.12). Thus at moderate temperatures, we have

$$E_{Fi} \geq E_{FP\ T>0K} \geq E_A$$

However, at moderate temperatures, it cannot touch the intrinsic Fermi level, because then it would wrongly indicate an equal concentration of electrons in conduction band and holes in valance band. Now if the temperature is extremely high, then the almost all the electrons in the valance band are excited to the conduction band. In such a case the concentration of intrinsic charge carriers overcomes the concentration of holes due to acceptor impurity. Thus the concentration of electrons in the conduction band and holes in the valance band becomes almost equal. Thus at extremely high temperatures, the P type semiconductor is converted in to intrinsic

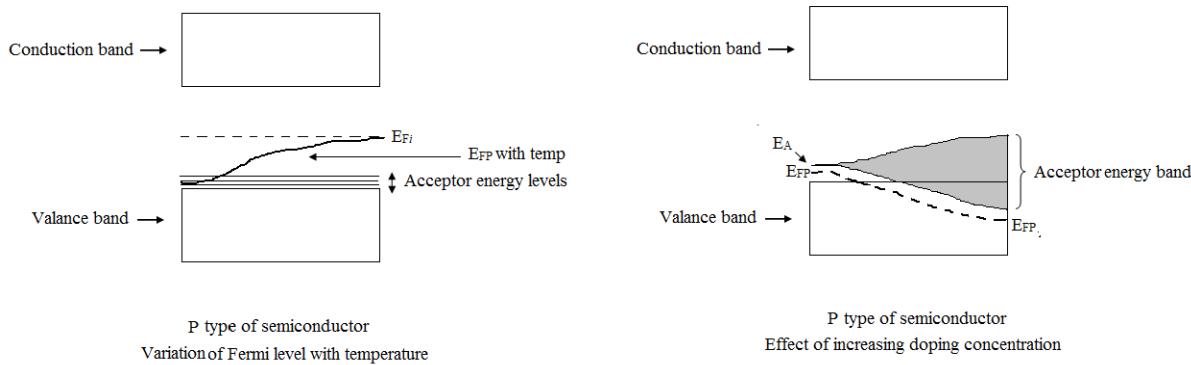


Fig (6.13): Effect of temperature and doping level on Fermi level in P type semiconductor

semiconductor. Thus now the Fermi level coincides with the intrinsic Fermi level. The variation of the P type Fermi level with temperature is shown in Fig. (6.13)

We know that in intrinsic semiconductor, doping level is kept very low. Thus the distance between the acceptor impurity atoms is comparatively large. Consequently the interaction between the acceptor impurity atoms is weak and thus the acceptor levels are discrete. However, when the doping level is increased, the atom-to-atom distance of the acceptor impurity decreases. Now the acceptor atoms interact and thus the acceptor levels split. When the doping level is further increased, the acceptor levels become acceptor band and then it penetrates in the valance band (Fig 6.13). As a result the P type Fermi level also penetrates the valance band.

Example (6.2) (Optional) : Find the probability with which an energy level 0.02 eV below Fermi level will be occupied at room temperature of 300 K and 1000 K. What do the results signify?

Solution:

We have

$$P(E) = \frac{1}{1 + e^{\frac{(E - E_F)}{kT}}}$$

As the energy level is below the Fermi level, $(E - E_F) = -0.02$

$$k = 1.38 \times 10^{-23} \frac{J}{K} = 8.63 \times 10^{-5} \frac{eV}{K}$$

Thus, at T = 300 K,

$$P(E) = \frac{1}{1 + e^{\frac{-0.02}{8.63 \times 10^{-5} \times 300}}} = 0.68 = 68\%$$

At T = 1000 K

$$P(E) = \frac{1}{1 + e^{\frac{-0.02}{8.63 \times 10^{-5} \times 1000}}} = 0.56 = 56\%$$

The results signify that, as temperature is increased, the probability of occupancy of the energy levels below the Fermi level decreases. This is because when temperature is increased, the electrons are excited from valence band to the conduction band. As a result, the energy levels below Fermi level are depleted leading to decrease in the probability of their occupancy.

Example (6.3) (Optional) :Find the probability of an electron occupying an energy level 0.02 eV above the Fermi level at 300 K and 1000 K. What do the results signify?

Solution:

We have

$$P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{kT}}}$$

As the energy level is above the Fermi level, $(E - E_F) = +0.02$

$$k = 1.38 \times 10^{-23} \frac{J}{K} = 8.63 \times 10^{-5} \frac{eV}{K}$$

Thus, at T = 300 K,

$$P(E) = \frac{1}{1 + e^{\frac{+0.02}{8.63 \times 10^{-5} \times 300}}} = 0.32 = 32\%$$

At T = 1000 K

$$P(E) = \frac{1}{1 + e^{\frac{+0.02}{8.63 \times 10^{-5} \times 1000}}} = 0.44 = 44\%$$

Above results lead to two implications

- i. As the temperature is increased, the probability of occupancy of the energy levels above the Fermi level increases. This is because, when temperature is increased, the electrons are excited. Such electrons occupy the energy levels above the Fermi level, as a result of which their probability of occupancy increases.
 - ii. It can also be noted that, the probability of occupancy of the energy levels above the Fermi level {P(E>E_F)} are exactly complementary to the probability of occupancy of energy levels below the Fermi level {P(E<E_F)}. This is because, when electron are excited the lower levels are depleted, and then the upper levels are occupied. This is further explained in the next example
-

Example (6.4) (Optional) : Show that if the probability of occupancy is x at the energy level ΔE below the Fermi level, then x is also probability of non-occupancy at an energy level ΔE above the Fermi level

Solution:

We know that when the energy level is below the Fermi level, ΔE is negative. Thus

$$P(E - \Delta E) = \frac{1}{1 + e^{\frac{(E-E_F)}{kT}}} \text{ becomes}$$

$$x = \frac{1}{1 + e^{\frac{-\Delta E}{kT}}} \quad \dots(6.12)$$

We also know that when the energy level is below the Fermi level, ΔE is positive. Thus probability of occupancy of energy level above the Fermi level can be written as

$$P(E + \Delta E) = \frac{1}{1 + e^{\frac{\Delta E}{kT}}}$$

Thus the probability of non-occupancy of energy level above the Fermi level is

$$1 - P(E + \Delta E) = 1 - \frac{1}{\left(1 + e^{\frac{\Delta E}{kT}}\right)}$$

$$= \frac{\left(1 + e^{\frac{\Delta E}{kT}}\right) - 1}{\left(1 + e^{\frac{\Delta E}{kT}}\right)}$$

$$= \frac{e^{\frac{\Delta E}{kT}}}{\left(1 + e^{\frac{\Delta E}{kT}}\right)}$$

Dividing both numerator and denominator by $e^{\frac{\Delta E}{kT}}$ we get

Probability of non – occupancy of the energy level above the Fermi level =

$$1 - P(E + \Delta E) = \frac{1}{\left(1 + e^{-\frac{\Delta E}{kT}}\right)} = x \text{ (according to Eqn. (6.12))}$$

Example (6.5) (Optional): At any given nonzero temperature, what is the probability of occupancy for a state whose energy is equal to Fermi energy?

Solution:

We have

$$P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{kT}}}$$

If $T \neq 0$ K and $E = E_F$, we have

$$P(E = E_F) = \frac{1}{1 + e^{\frac{(E_F-E_F)}{kT}}} = 0.5 = 50\%$$

This result is consistent with our previous discussion. Note that the result is same at any value of nonzero temperature.

Example (6.6) (Optional) : Find the temperature at which there is 1 % probability that a state with energy 0.5 eV above Fermi energy will be occupied. What does the result signify?

Solution:

$$P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{kT}}}$$

Substituting required values

$$\begin{aligned} 0.01 &= \frac{1}{1 + e^{\frac{5793}{T}}} \\ \Rightarrow \frac{1}{0.01} &= \frac{1 + e^{\frac{5793}{T}}}{1} \\ \Rightarrow 100 &= 1 + e^{\frac{5793}{T}} \\ \Rightarrow 99 &= e^{\frac{5793}{T}} \\ \Rightarrow \ln 99 &= \frac{5793}{T} \\ \Rightarrow 4.595 &= \frac{5793}{T} \Rightarrow T = 1261 \text{ K} \end{aligned}$$

The result indicate that temperature needs to be considerably high, even for achieving 1% probability of occupancy of a level 0.5 above the Fermi level. This also signifies that when the energy level is high, it is more difficult to excite the electron there. We can also note that, for a given temperature, as we move gradually above the Fermi level, the probability of occupancy decreases.

6.9 QUANTITATIVE DESCRIPTION OF SEMICONDUCTORS(Derivations are optional, but highlighted formulae are compulsory):

How to characterize the semiconductors

Till now we have seen that the behavior of a semiconductor in various conditions such as temperature or doping level can be described appropriately by means of Fermi level. Quantitative description of semiconductors is also necessary, especially for measuring their properties. Various physical quantities such as carrier concentration (n), drift velocity (v_d), mobility (μ), current (I), current density (J), resistance (R), resistivity (ρ), conductivity (σ), voltage (V), electric field (E) etc. are associated with the behavior of semiconductors in various conditions. Thus it is of interest to understand these physical quantities and relations between them.

At thermal equilibrium and in absence of electric field, the current in intrinsic as well as extrinsic semiconductors is zero. Else, when equilibrium is disturbed or when electric field is applied, there occurs a motion of the charge carriers. The current which occurs due to a gradual drift of electrons towards the positive terminal (or holes towards the negative terminal) is called as **drift current**. The corresponding velocity is called as **drift velocity (v_d)**. Note that while moving through the crystals, the charge carriers will suffer collisions with the ions or lattice points. Thus the drift motion is random (Fig). Thus though the charge carrier, say electron, moves ultimately towards the positive polarity, it is scattered several times in various directions due to the collisions in between. Therefore the ultimate drift velocity is very less as compared to the instantaneous velocity (the velocity between two successive collisions). Thus we may define drift velocity as

$$v_d = \frac{\Delta l}{\Delta t} \text{ (by taking collisions into consideration)}$$

Typically drift velocities are $\sim 10^{-5}$ m/s. Drift velocity depends upon the applied electric field. A quantity which is more fundamental and independent of the electric field is **mobility**. It is defined as drift velocity acquired per unit electric field

$$\mu = \frac{v_d}{E} \quad \dots(6.13)$$

Mobility thus measures the easiness by which charge carriers move through the specimen. The unit of mobility is

$$\frac{m/s}{V/m} = \frac{m^2}{V.s}$$

The values of mobility depend upon the type of specimen, temperature as well as doping level. Typical values of mobility of electrons and holes in silicon at room temperature is $1400 \text{ cm}^2/\text{V.s}$ and $450 \text{ cm}^2/\text{V.s}$

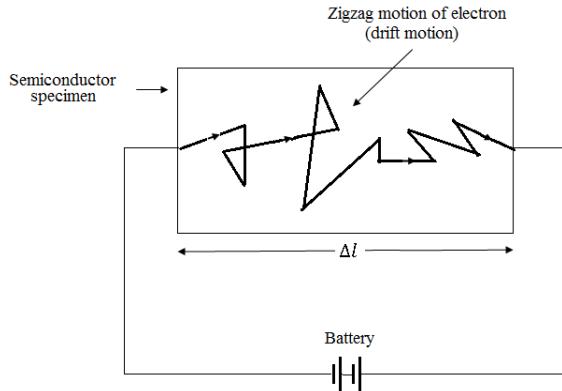


Figure (6.14) Drift motion

Let us now discuss the other physical quantities and their interrelations. We know that current is the rate of flow of charge. Thus

$$I = \frac{\Delta Q}{\Delta t} \quad \dots(6.14)$$

The net charge depends upon number of electrons (or holes), thus

$$I = \frac{Nq}{\Delta t}$$

Where, N is the total number of electrons (or holes) and q is the magnitude of their charge. Now let us divide and multiply above expression by volume (V) of the specimen.

$$I = \frac{Nq}{\Delta t} \times \frac{V}{V}$$

Rearranging and recalling that volume is the product of area (A) and length(Δl),

$$I = q \left(\frac{N}{V} \right) \left(\frac{\Delta l}{\Delta t} \right) A$$

The quantity $\left(\frac{N}{V} \right)$ represents the carrier concentration/charge particle density/number of charge carriers per unit volume (denoted by n). The quantity $\left(\frac{\Delta l}{\Delta t} \right)$ represents the drift velocity (v_d). Thus we have

$$I = nqv_d A \quad \dots(6.15)$$

Eqn (6.15) is an important expression which relates the current with the properties of the current carrying specimen (n , q , v_d and A). It is possible to express Eqn (6.15) in terms of current density and mobility.

Recalling that $J = \frac{I}{A}$ and $\mu = \frac{v_d}{E}$, we have

$$J = \frac{I}{A} = nqv_d \quad \dots(6.16)$$

$$J = \frac{I}{A} = nq\mu E \quad \dots(6.17)$$

Now let us recall Ohm's law

$$V = IR \quad \dots(6.18)$$

Dividing both the sides by Δl and multiplying by A , we get

$$\begin{aligned} \frac{V}{\Delta l} A &= \frac{IAR}{\Delta l} \\ \Rightarrow \frac{V}{\Delta l} &= \frac{I}{A} \frac{R}{\Delta l/A} \end{aligned}$$

Recalling that electric field, $E = \frac{V}{\Delta l}$, current density, $J = \frac{I}{A}$ and resistivity, $\rho = \frac{R}{\Delta l/A}$, we get

$$E = J\rho \quad \dots(6.19)$$

Above Eqn expresses Ohm's law in terms of E , J and ρ . Note that the quantities V , I and R change with dimensions of the specimen however the quantities E , J and ρ do not depend upon the dimensions. They depend only upon material. (for ex more current will flow through a specimen with larger cross sectional area, however for a given specimen, the current density will remain same, irrespective of the area) The expression $E = J\rho$ can also be expressed as

$$J = \frac{E}{\rho} = E\sigma \quad \dots(6.20)$$

Comparing Eqns (6.19) and (6.20), we get

$$\sigma = nq\mu \quad \dots(6.21)$$

Eq. (6.21) is a fundamental expression which expresses conductivity in terms of carrier concentration, and mobility.

It can be seen that Eqn (6.15) and (6.21) involve n , the carrier concentration. As metals contain only one kind of charge carrier (electron), we can use these Eqns as it is for metals. However, semiconductors contain two kinds of charge carriers; electrons and holes. Thus, for semiconductors, we must write

$$\sigma = n_e e \mu_e + n_h e \mu_h \quad \dots(6.22)$$

Where e represents the charge on electron (as well as hole), n_e and n_h represents carrier concentration of electrons and holes and μ_e and μ_h represent mobility of electrons and holes respectively. Now recalling that the carrier concentrations of electrons and holes in intrinsic semiconductors are equal, we have $n_e = n_h = n_i$. Thus

$$\sigma_i = n_i e (\mu_e + \mu_h) \quad \dots(6.23)$$

Note that the mobility of electrons and holes are not equal. Indeed holes have less mobility as compared to electrons.

We know that for N type semiconductors $n_e \gg n_h$ and for P type semiconductors $n_h \gg n_e$. Thus we have

$$\sigma_N = n_e e \mu_e \quad \text{for N type semiconductors and}$$

$$\sigma_P = n_h e \mu_h \quad \text{for P type of semiconductors}$$

On similar lines

$$I_i = e n_i A (\mu_e + \mu_h) E \quad \text{for intrinsic semiconductors}$$

$$I_N = e n_e \mu_e A E \quad \text{for N type semiconductors and}$$

$$I_P = e n_h \mu_h A E \quad \text{for P type semiconductors}$$

Note that, in the expressions for the conductivity, temperature is not included. Although, the mobility hardly depends upon temperature, the carrier concentration and hence the conductivity strongly depends upon the temperature.

It can be shown that

$$n_e = 2 \left[\frac{2\pi m_e^* k T}{h^2} \right]^{\frac{3}{2}} e^{-\frac{(E_C - E_F)}{kT}} \quad \dots(6.24)$$

$$n_h = 2 \left[\frac{2\pi m_h^* k T}{h^2} \right]^{\frac{3}{2}} e^{-\frac{(E_F - E_V)}{kT}} \quad \dots(6.25)$$

The intrinsic carrier concentration can be shown to be equal to

$$n_i = 2 \left[\frac{2\pi kT}{h^2} \right]^{\frac{3}{2}} (m_e^* m_h^*)^{\frac{3}{4}} e^{-\frac{E_g}{2kT}} \quad \dots(6.26)$$

Thus, it can be observed that the carrier concentration increases with temperature. Substituting n_i from above expression in (6.26), it can be shown that

$$\sigma = \sigma_0 e^{-\frac{E_g}{2kT}} \quad \dots(6.27)$$

It can be noted that, as temperature increases, the conductivity decreases in an exponential manner. This is due to the thermal excitation of the electrons in the conduction band. It can also be seen that, as band gap (E_g) increases the conductivity decreases.

All the expressions that we have obtained till now help us to make quantitative estimates of the properties of the semiconductors

Example (6.6). Calculate the free electron density in copper, if each copper atom donates one electron to the conduction band. (Properties of copper: Density = 8.96 gm/cc, atomic weight 63.5 and Avogadro's number = 6.02×10^{23} atoms/mole)

Solution: We know that

$$\begin{aligned} & 63.5 \text{ gm of copper contains } 6.023 \times 10^{23} \text{ atoms} \\ \Rightarrow & 1 \text{ gm of copper contains } \frac{6.023 \times 10^{23} \text{ atoms}}{63.5 \text{ gm}} \\ \Rightarrow & 8.96 \frac{\text{gm}}{\text{cc}} \text{ of copper contains } \frac{6.023 \times 10^{23}}{63.5} \left(\frac{\text{atoms}}{\text{gm}} \right) \times 8.96 \left(\frac{\text{gm}}{\text{cc}} \right) \\ & = 8.5 \times 10^{22} \frac{\text{atoms}}{\text{cc}} \\ & = 8.5 \times 10^{22} \frac{\text{electrons}}{\text{cc}} \quad \text{as each copper atom donates 1} \\ & \quad \text{electron in conduction band} \end{aligned}$$

Example (6.7): Calculate the free electron density in silicon, if 3.33×10^{12} atoms of silicon donate one electron in conduction band (Properties of silicon: Density = 2.3290 g/cm^3 , Atomic weight = 28.085, Avogadro's number = 6.02×10^{23}). Also calculate the density of holes in the valance band

Solution:

28.085 g of silicon contain 6.023×10^{23} atoms/mole

$$\Rightarrow 1 \text{ gm of silicon contain } \frac{6.023 \times 10^{23} \text{ atoms/mole}}{28.05 \text{ g/mole}} = 2.15 \times 10^{22} \frac{\text{atoms}}{\text{g}}$$

$$\Rightarrow 2.3290 \frac{\text{g}}{\text{cc}} \text{ of silicon contain } 2.15 \times 10^{22} \frac{\text{atoms}}{\text{g}} \times 2.3290 \frac{\text{g}}{\text{cc}} = 5 \times 10^{22} \frac{\text{atoms}}{\text{cc}}$$

As per the given data, 3.33×10^{12} atoms of silicon donate one electron in conduction band

$$\text{Thus the number density of electrons is } \frac{5 \times 10^{22}}{3.33 \times 10^{12}} = 1.5 \times 10^{10} \frac{\text{electrons}}{\text{cc}}$$

As the semiconductor is intrinsic, same will be the density of holes in the valance band

It can be seen that the charge carrier concentration of the electrons in intrinsic silicon is significantly less as compared to copper. This is due to their band gaps. As discussed earlier, the charge carrier concentration (and hence the conductivity) can be enhanced by doping.

Example (6.8): Resistance of copper wire of diameter 1.03 mm is 6.51 ohm per 300 m. The concentration of free electrons in copper is $8.4 \times 10^{28} / \text{m}^3$. If the current is 2 A, find the mobility of free electrons

Solution:

This problem does not require the value of current

$$\text{Diameter of copper wire} = 1.03 \text{ mm} = 1.03 \times 10^{-3} \text{ m}$$

$$\text{Radius} = \frac{1.03 \times 10^{-3}}{2} = 5.15 \times 10^{-4} \text{ m}$$

$$\text{Area } \pi r^2 = 3.14 \times (5.15 \times 10^{-4})^2 = 8.33 \times 10^{-7} \text{ m}^2$$

$$\text{We know that } \rho = \frac{RA}{l} = \frac{6.51 \times 8.33 \times 10^{-7}}{300} = 1.81 \times 10^{-8} \Omega \cdot \text{m}$$

$$\sigma = \frac{1}{\rho} = \frac{1}{1.81 \times 10^{-8}} = 5.5 \times 10^7 \frac{\text{mho}}{\text{m}}$$

We have

$$\sigma = ne\mu$$

$$\Rightarrow \mu = \frac{\sigma}{ne} = \frac{5.5 \times 10^7}{8.4 \times 10^{28} \times 1.6 \times 10^{-19}} = 4.11 \times 10^{-3} \frac{\text{m}^2}{\text{V.s}}$$

We can note that this mobility is quite less. However, copper is the best conductor of electricity. This is because the concentration of free electrons is very high

Example (6.9): Calculate the conductivity of pure/intrinsic silicon if free electron concentration is $1.5 \times 10^{10} /cm^3$ and mobility of electrons and holes are $1500 \text{ cm}^2/\text{V.s}$ and $500 \text{ cm}^2/\text{V.s}$ respectively

Solution:

We have

$$\sigma_i = en_i(\mu_e + \mu_h)$$

Given

$$n_i = 1.5 \times 10^{10} /cm^3 = 1.5 \times 10^{16} /m^3$$

$$\mu_e = 1500 \frac{\text{cm}^2}{\text{V.s}} = 0.15 \frac{\text{m}^2}{\text{V.s}} \text{ and } \mu_h = 500 \frac{\text{cm}^2}{\text{V.s}} = 0.05 \frac{\text{m}^2}{\text{V.s}}$$

Substituting

$$\sigma_i = 1.6 \times 10^{-19} \times 1.5 \times 10^{16} \times (0.15 + 0.05) = 4.8 \times 10^{-4} \text{ mho/m}$$

As it can be seen, this conductivity is quite less as compared to conductors. This because the charge carrier concentration of intrinsic silicon is considerably less as compared to the conductors. Conductivity of silicon can be enhanced by doping

Example (6.10): Intrinsic silicon is doped with phosphorus with the atomic ratio of 10^8 (Si) : 1 (P). Calculate the conductivity of N type of silicon thus formed. (Properties of silicon as given in examples ... and Given: mobility of electrons in N type silicon = $0.036 \frac{\text{m}^2}{\text{V.s}}$. Compare this conductivity with that of intrinsic silicon and comment on the result.

Solution: By referring example (), we have

$$\text{Number density of silicon atoms} = 5 \times 10^{22} \frac{\text{atoms}}{\text{cc}}$$

10^8 silicon atoms contain 1 phosphorous atom

$$\Rightarrow 5 \times 10^{22} \frac{\text{atoms}}{\text{cc}} \text{ of silicon contain } \frac{5 \times 10^{22}}{10^8} = 5 \times 10^{14} \text{ phosphorous atoms/cc}$$

Each phosphorous atom donates 1 electron in conduction band

$$\text{Thus free electron density of N type silicon} = n_e = 5 \times 10^{14} \frac{\text{electrons}}{\text{cc}} = 5 \times 10^{20} \frac{\text{electrons}}{\text{m}^3}$$

We have

$$\text{conductivity} = \sigma_N = n_e e \mu_e$$

$$\Rightarrow \sigma_N = 5 \times 10^{20} \times 1.6 \times 10^{-19} \times 0.036$$

$$\Rightarrow \sigma_N = 2.88 \frac{mho}{m}$$

When compared to the conductivity of intrinsic silicon in example () ($4.8 \times 10^{-4} mho/m$), we find that the conductivity of N type of silicon is 6 thousand times better than intrinsic silicon. Thus we find that, a very small percentage of dopant (1 dopant atom per 100 million silicon atoms) can bring about a significant enhancement in the conductivity. Note that a still higher concentration of dopant may bring about a still better improvement in the conductivity. However, the dopant percentage is deliberately kept low, to keep the crystal structure of the host intact.

We may also note that, when compared to the conductivity of copper in example () ($5.5 \times 10^7 \frac{mho}{m}$), the conductivity of N type of silicon is yet considerably (10^{-8} times less). This is mainly due to high carrier concentration in copper.

	Edwin Herbert Hall (1855 – 1938): He was an American Physicist. He obtained degree in Johns Hopkins University. He was later appointed as Professor of Physics in Harvard university. While doing experimental work during Ph.D, he discovered (at the age of 24) that when a current carrying semiconductor (and even some conductors) is exposed to a transverse magnetic field, a voltage appears across its faces. This effect now called Hall effect was published by him in a reputed American journal. As we shall see it further, Hall effect has some important practical applications. Two special cases of Hall effect have also been discovered. These are Anomalous Hall effect and Quantum Hall effect (Klaus von Klitzing: Nobel prize in Physics in 1985).
--	---

6.10 HALL EFFECT:

How semiconductors behave in magnetic field

Till now, we have discussed both N and P type of semiconductors. N type semiconductor mainly conducts due to electrons and P type of semiconductor mainly conducts due to holes. We also know that hole behaves like a positive charge carrier. Is there any experimental method to identify whether a given specimen is of N or P type? Is it possible to identify whether a specimen is conducting due to negative or positive kind of charge carriers? Hall effect makes this possible.

We know that motion of a charge carrier in vacuum is affected due to magnetic field. This is true even if the charge carrier drifts through a solid (say a conductor or semiconductor). We know that conductors and N type semiconductors conduct due to electrons. Consider a specimen conducting due to electrons (Fig 6.16). Due to externally applied electric field (E_x), the electron drifts towards the positive polarity (in +X direction). Now if this current carrying specimen is kept in a transverse magnetic field (+Z direction, going inside the plane of paper), then, the electron experiences a Lorentz force acting in +Y direction. This force is given by

$$F_B = q\vec{v}_d \times \vec{B} \quad \dots(6.28)$$

The charge on the electron is $-e$. As v_d and B are acting in $+X$ and $+Z$ (mutually perpendicular) directions, the angle θ is 90° . Thus we get

$$F_B = -ev_d B \sin\theta = -ev_d B \quad \dots(6.29)$$

Now right hand thumb rule indicates that F_B will be perpendicular to both v_d and B , that is in $+Y$ direction. However, due to negative sign of the electronic charge it will act in $-Y$ direction. This force will deflect the electrons in downward direction. Thus now electrons will not follow the external electric field, and instead of moving in $+X$ direction, they will move downwards. Thus gradually, the electrons will pile up the lower surface. As many electrons in the entire specimen move downward, the upper surface will lose electrons and thus acquire an uncompensated positive charge (Fig. 6.16).

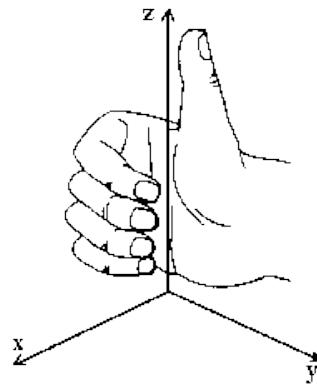


Figure (6.15): Right hand thumb rule

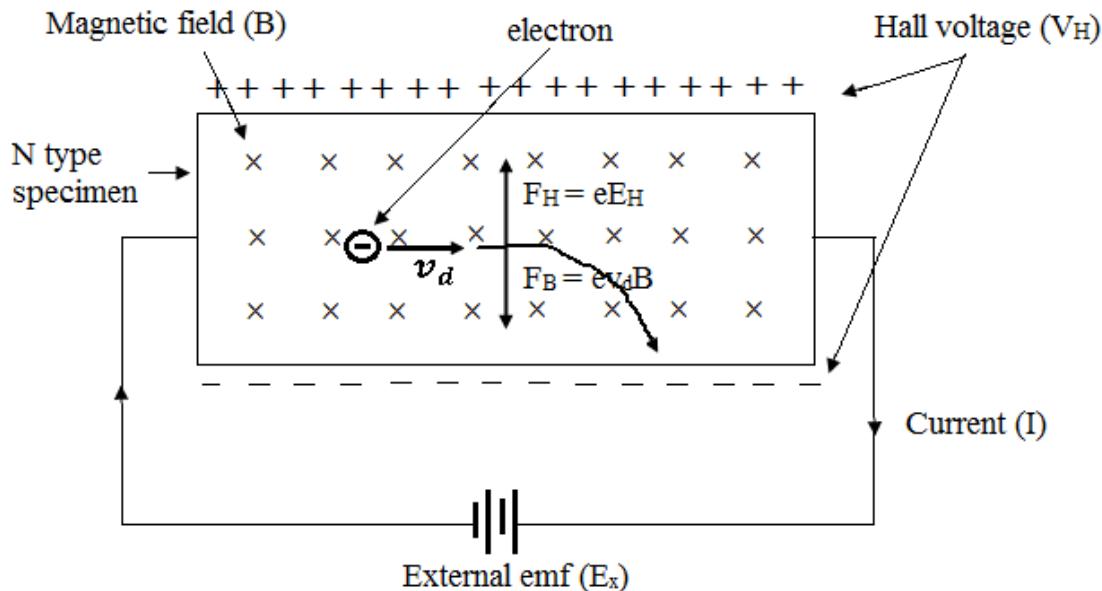


Figure (6.16): Hall effect in N type semiconductor

Thus a p.d. will be developed across the width of the specimen. This p.d. is called Hall voltage (V_H). The corresponding Hall electric field is given by

$$E_H = \frac{V_H}{w} \dots (6.30)$$

This electric field will act in downward that is $-Y$ direction. As this electric field gradually grows, it develops a Hall force given by

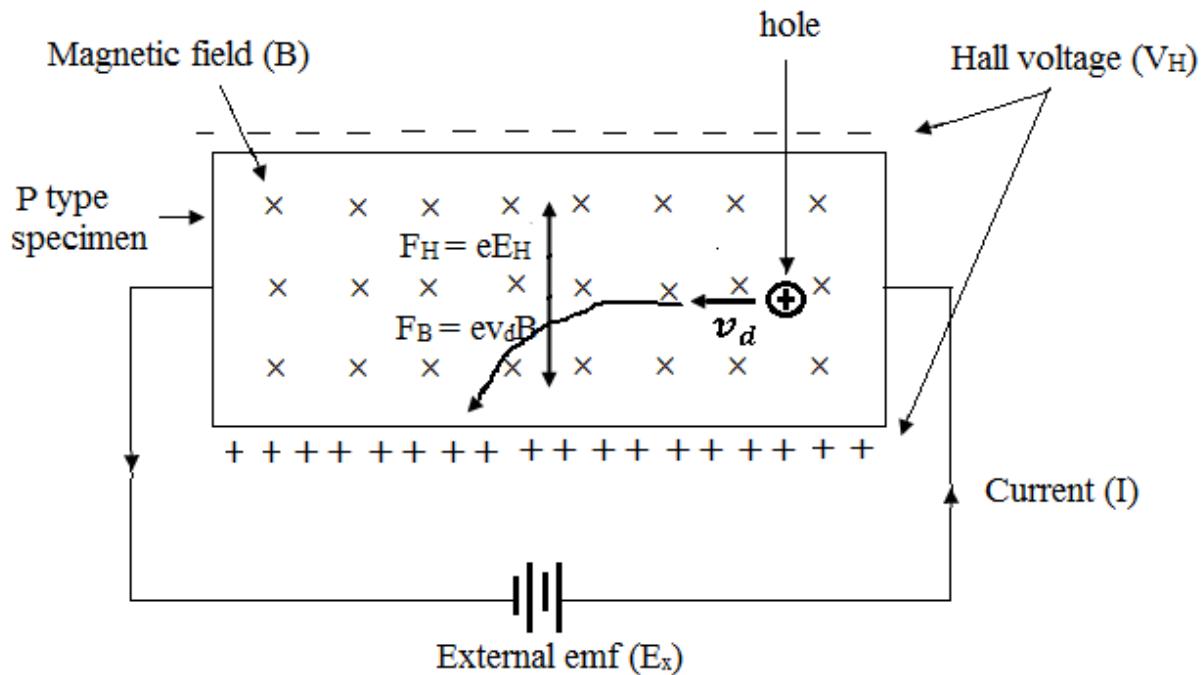


Figure (6.17): Hall effect in P type semiconductor

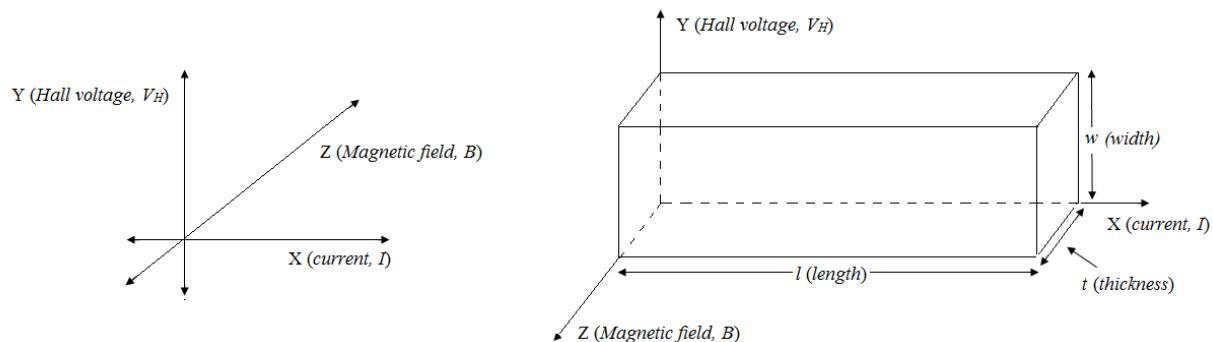


Figure (6.18) Relative directions of current, magnetic field and Hall voltage

$$F_H = -e\vec{E}_H \quad \dots(6.31)$$

This acts in + Y direction and thus starts opposing the further magnetic deflection of electrons in -Y direction. Due to piling up of electrons, the strength of Hall force gradually increases. Note that Hall force (F_H) acts in +Y direction and the magnetic force (F_B) acts in -Y direction. At a given stage, the Hall force compensates the magnetic force. (Fig 6.16) and the net force along Y direction becomes zero. At this stage, the downward deflection of electrons ceases and thus now electrons continue to follow the external electric field. This situation is the equilibrium situation. The ceasing of electrons moving in downward direction can also be explained by considering that the lower surface is negative and thus should oppose the downward motion of electrons. Secondly, the Hall electric field is acting in downward direction and electrons cannot move along the electric field. Thus their -Y motion is opposed. In equilibrium condition, we can equate the magnitudes of downward magnetic force and upward Hall force. Thus we get

$$\begin{aligned} F_H &= F_B \\ \Rightarrow eE_H &= ev_d B \\ \Rightarrow \frac{V_H}{w} &= v_d B \end{aligned} \quad \dots(6.32)$$

From Eqn (6.15), we have

$$\begin{aligned} I &= qnv_d A \\ \Rightarrow v_d &= \frac{I}{nqA} \end{aligned}$$

Substituting v_d in Eqn (6.32)

$$\frac{V_H}{w} = \frac{I}{nqA} B$$

Rearranging

$$V_H = \frac{1}{nq} IB \frac{w}{A} \quad \dots(6.33)$$

As $J = \frac{I}{A}$ we can also write

$$V_H = \frac{1}{nq} JBw \quad \dots(6.34)$$

From Fig (6.18)

$$\text{crosssectional area} = \text{width} \times \text{thickness}$$

$$\Rightarrow A = wt$$

Substituting in Eqn (6.33)

$$V_H = \frac{1}{nq} IB \frac{w}{wt}$$

$$\Rightarrow V_H = \frac{1}{nq} IB \frac{1}{t} \quad \dots(6.35)$$

Eqns (6.34) and (6.35) give standard expression for Hall voltage. It can be noted that Hall voltage increases by increasing current and magnetic field. However it decreases if carrier concentration is increased. This can be interpreted by considering that increase in the carrier concentration will result in more number of collisions, causing decrease in the drift velocity. Hall voltage directly depends upon drift velocity Eqn (6.32). Thus Hall voltage is inversely proportional to carrier concentration. This indicates why conductors are less sensitive to the Hall Effect than the semiconductors. The carrier concentration of semiconductors is 1 lakh times smaller than conductors. Thus Hall voltage in semiconductors is 1 lakh times greater than semiconductors.

Above discussion applies to electrons, i.e. N type semiconductors (and metals). Now let us consider a P type semiconductor. Refer Fig (6.17). Note that all experimental conditions are kept same, except that N type of semiconductor is replaced by P type semiconductor. We know that Ptype semiconductor conducts due to holes which behave like a positive charge carrier. Thus now holes will drift in opposite i.e. in $-X$ direction. Recall that magnetic field is acting in $+Z$ direction (going inside the plane of paper). Right hand thumb rule indicates that even in this case the magnetic force (in Eqn (6.28)) will deflect the positive holes in downward direction. Thus the holes will pile up on the lower surface. (Note that in previous case electrons piled up on the lower surface). As the entire specimen consists of holes, all holes will start moving downwards. The upper surface will thus become devoid of holes and thus will acquire an uncompensated negative charge. Thus, a Hall voltage develops in this case also, but as it can be seen, the Hall voltage for P type specimen is exactly opposite to the N type semiconductor. This clearly indicates that Hall effect helps to identify the type (N or P) of semiconductor. If a voltmeter is connected across the width, the deflection produced in the voltmeter for N type semiconductor is exactly opposite to P type semiconductor. Note that the experimental conditions i.e. directions of magnetic field and external battery are kept same.

Note that in the case of P type semiconductor, the direction of the Hall voltage, Hall electric field is opposite to that in N type semiconductor. Thus Hall electric field is directed upwards. Holes, which behave like a positive charge carrier, cannot move opposite to the direction of electric field. Thus the Hall force on the holes is directed upwards and it gradually compensates the magnetic force in $-Y$ direction. At equilibrium, Hall force and magnetic force are equal and opposite. Thus when equilibrium occurs, the holes continue to move in $-X$ direction. This also indicates that Eqn. (6.35), which has been derived for electrons (N type semiconductor) are equally applicable to holes (P type semiconductor) except that sign of the charge and the direction of Hall voltage are opposite.

Now let us find out the significance of the term $\frac{1}{nq}$ in Eq (6.35). We note that for a given specimen, the Hall voltage is directly proportional to I, B and $\frac{w}{A}$. The factor $\frac{1}{nq}$ in Eqn (6.35) thus acts as a constant of proportionality between V_H and I, B, $\frac{w}{A}$. Thus in the given experimental condition, the strength of the Hall voltage depends upon $\frac{1}{nq}$. The quantity $\frac{1}{nq}$ is called as **Hall coefficient**. It can be noted that the Hall coefficient for semiconductors is greater than that for conductors. Indeed, the sensitivity of the material to Hall effect is governed by its Hall coefficient. We also note that the Hall coefficient for N type semiconductors (and metals) is negative while it is positive for P type semiconductors. As intrinsic semiconductor contains equal concentration of electrons and holes, the Hall voltage for intrinsic semiconductors is zero. Indeed, Hall effect is absent in intrinsic semiconductors.

We note that

$$\text{Hall coefficient} = R_H = \frac{1}{nq} \quad \dots (6.36)$$

As n in the above Eqn. represents carrier concentration $\left(\frac{1}{m^3}\right)$, the SI unit of Hall coefficient is $\left(\frac{m^3}{C}\right)$.

Table (6.8) shows the Hall coefficients of some elements. Note that for beryllium, cadmium, zinc and antimony, the Hall coefficients are positive indicating that holes exist in some metals also. Indeed this was the first experimental evidence for the existence of holes.

Element	$R_H \times 10^{-10} \text{ m}^3/\text{C}$
Na	-2.5
K	-4.2
Cu	-0.55
Ag	-0.84
Al	-0.30
Au	-0.71
Be	2.4
Zn	0.33
Cd	0.6
Sb	230
Co	-1.33
Bi	-100
Ni	-0.61
Pt	-0.23

Table (6.8): Hall coefficient of some elements. Note that in some metals it is positive

Eq (6.35) indicates that if current and magnetic field are known and if Hall voltage is measured then Hall coefficient can be estimated. Knowledge of Hall coefficient provides an estimate of charge carrier concentration. As carrier concentration in semiconductors depends upon doping level, Hall effect also provides a method to measure doping level. We can also note that, if we use a specimen with known Hall coefficient and if current is known and if Hall voltage is measured, then magnetic field can also be measured. Such devices are called as **Hall probes (gauss-meter)**. The advantage of Hall probe for measuring the magnetic field is that, both weaker and stronger magnetic fields can be measured and direction of the magnetic field can also be identified. Further Hall probe can be used to measure steady as well as alternating magnetic fields. High frequency alternating magnetic fields (up to 10^{12} Hz) can also be measured.

We know that

$$\sigma = nq\mu$$

$$\Rightarrow \sigma = \frac{1}{R_H} \mu$$

$$\Rightarrow \mu = \sigma R_H \quad \dots(6.37)$$

Thus Hall effect also provides a method to measure mobility (provided that conductivity is known). In addition to mobility, Hall effect also provides an effective method of measuring the drift velocity. In this case the specimen is moved with varying speeds till the Hall voltage becomes zero. Note that if the specimen is moved in opposite direction with a velocity exactly equal to drift velocity, the resultant velocity reduces to zero and thus Hall voltage becomes zero.

Now let us introduce an additional physical quantity related to Hall effect. Due to occurrence of Hall effect, the charge carriers experience two electric fields, the external electric field (E_x) acting in X direction and the Hall electric field (E_H) acting in Y direction. The resultant electric field will thus act along an angle θ_H given by

$$\tan \theta_H = \frac{E_H}{E_x}$$

θ_H is called as **Hall angle**. It can be shown that

$$\tan \theta_H = \mu B$$

Let us now summarize the applications of Hall effect

- i. Identifying the type of charge carrier (positive or negative)
- ii. Identifying the type of semiconductor (P or N)
- iii. Measuring charge carrier concentration
- iv. Measuring doping level
- v. Measuring the mobility of charge carriers

- vi. Measuring magnetic field and its direction with accuracy (steady as well as alternating, weak as well as strong)
 - vii. Measuring the current (weak as well as heavy) in a cable without interrupting the circuit. (We know that current in a conductor generates magnetic field around the cable. If this magnetic field is measured then the current can be estimated. This can be done without breaking the wire)
-

Example (6.11): A strip of copper having thickness 0.5 mm is placed in a magnetic field of magnitude 0.75 T. A current of 100 mA is sent through the strip. What is the Hall potential difference that will appear across the width of the strip? The carrier concentration of electrons in copper is 8.47×10^{28} electrons/m³.

Solution:

We have $V_H = \frac{1}{nq} BI \frac{1}{t}$

Thus

$$V_H = \frac{1}{1.6 \times 10^{-19} \times 8.47 \times 10^{28}} \times 0.75 \times 100 \times 10^{-3} \times \frac{1}{0.5 \times 10^{-3}}$$

$$V_H = 11 \times 10^{-12} V = 11 \text{ pV}$$

The result indicates that the Hall voltage is very weak. This is consistent with the fact that conductors are very weakly sensitive to Hall effect.

Example (6.12): A copper specimen having length 1 m, width 1 cm, and thickness 1 mm is conducting with 1 A along its length and is applied with a magnetic field of 1 T along its thickness. It experiences Hall effect and a Hall voltage 0.074 μV appears along its thickness. Calculate the Hall coefficient, electron concentration and the mobility of electrons in copper. Conductivity of copper is $5.8 \times 10^7 (\Omega\text{m})^{-1}$

Solution:

We have

$$\begin{aligned} V_H &= \frac{1}{nq} IB \frac{w}{A} = R_H IB \frac{w}{A} \\ \Rightarrow 0.074 \times 10^{-6} &= R_H \times 1 \times 1 \times \frac{1 \times 10^{-2}}{1 \times 10^{-2} \times 1 \times 10^{-3}} \\ \Rightarrow R_H &= 7.4 \times 10^{-11} \frac{m^3}{C} \end{aligned}$$

We have

$$R_H = \frac{1}{nq}$$

$$\Rightarrow 7.4 \times 10^{-11} = \frac{1}{n \times 1.6 \times 10^{-19}}$$

$$\Rightarrow \text{carrier concentration} = n = 8.45 \times 10^{28} \frac{\text{electrons}}{\text{m}^3}$$

We have

$$\mu = \sigma R_H$$

$$\Rightarrow \mu = 5.8 \times 10^7 \times 7.4 \times 10^{-11}$$

$$\Rightarrow \mu = 4.3 \times 10^{-3} \frac{m^2}{V.s}$$

Example (6.13): N type semiconductor having length 1.00 cm, width 1.00 mm and thickness 0.1 mm is made to conduct with 1 mA current and is placed in the magnetic field acting along its thickness. The Hall voltage is measured to be 3.44×10^{-7} V. Calculate the magnetic field, if the Hall coefficient of the specimen is $-3.44 \times 10^{-8} \frac{m^3}{C}$

Solution:

We have

$$V_H = \frac{1}{nq} IB \frac{w}{A} = R_H IB \frac{w}{A} = R_H IB \frac{1}{t}$$

Thus

$$3.44 \times 10^{-7} = 3.44 \times 10^{-8} \times 1 \times 10^{-3} \times B \times \frac{1}{0.1 \times 10^{-3}}$$

$$\Rightarrow B = 1 T$$



Russell Shoemaker Ohl (1898 – 1987) was an American Physicist, who invented the PN junction diode in 1939. Ohl was a notable semiconductor researcher. His investigations on certain materials, which he carried out in AT&T's Bell Labs, resulted in invention of diode detectors suitable for high-frequency wireless, broadcasting, and military radar. Ohl, in 1939, invented P–N junction. He also established the concepts of barrier potential, and the unidirectional conductivity of diodes. He was the first to use super-purified germanium for making diodes. All diodes (including LEDs, laser diodes, photodiodes, solar cells etc.) are descendants of Ohl's work. His work with diodes led him later to develop the first silicon solar cells for which he received a patent.

6.11 PN JUNCTION DIODE:

Fermi level explains why diode conducts only in one direction

In microelectronics, millions of electronic components such as diodes, transistors, resistors and capacitors are made on tiny integrated circuits. Out of these, the PN junction is the simplest one. When P and N types of semiconductor are joined, a PN junction diode is formed. One of the methods to make such PN junctions is to diffuse pentavalent and trivalent impurity from opposite sides in to a pure silicon wafer. PN junction diode is simplest electronic device that can be made using P and N types of semiconductors. It is simplest, and yet one the most useful electronic device. PN junction diodes have variety of applications that include rectifier, voltage regulator (Zener), varactor (variable capacitor), varistor (variable resistor), tunnel diode, photodiode, light emitting diode (LED), diode-laser, sensors (of temperature and light), switch and solar cell. A PN junction is an inbuilt component of several electronic devices and circuits. It thus necessary to understand working of PN junction diode

Unbiased diode just at the moment of formation

Consider a PN junction diode just at the moment of its formation. Such diode is in non-equilibrium condition. The P side contains majority holes and minority electrons while N side contains majority electrons and minority holes. The concentration of electrons in N region is quite high as compared to P region. Thus there is a concentration gradient of electrons from N to P side. As discussed earlier, in the energy band diagram, the Fermi level on P side is close to valance band while on N side it is close to conduction band. As Fermi level behaves like a center of gravity, the closeness of the Fermi level to the conduction band on N side indicates that the

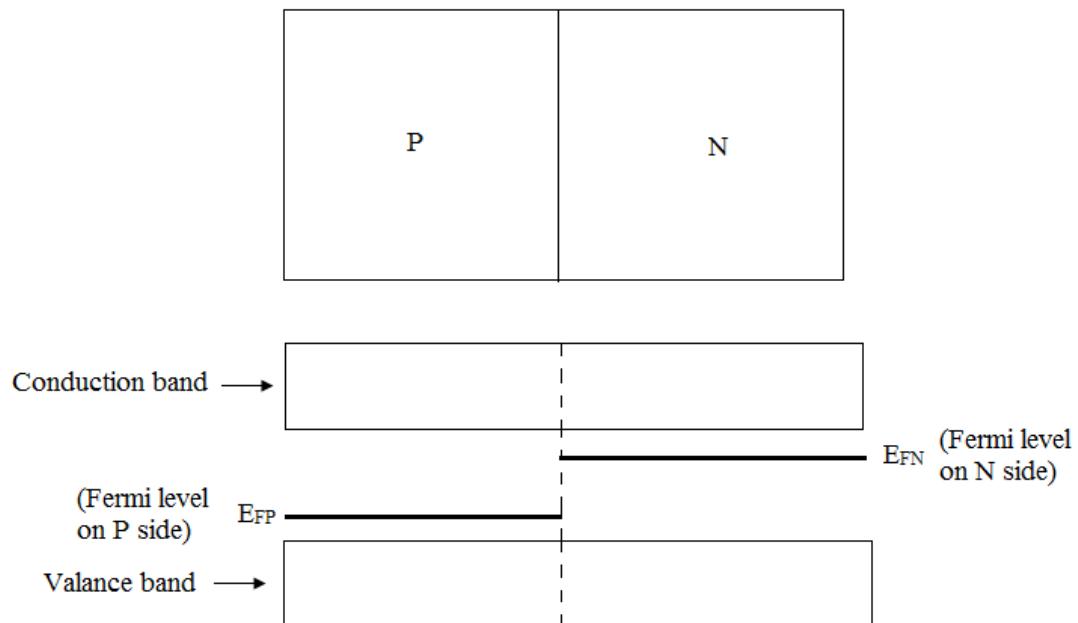


Figure (6.19): PN junction diode just at the moment of formation

electrons on N side are in majority and their overall energy is high. The closeness of Fermi level to the valance band on P side indicates that the lower energy states of the electrons are empty. This is a non-equilibrium situation.

Unbiased diode a few moments after its formation:

To establish the equilibrium, the concentration of the electrons on N as well as P sides should become equal. Similarly, the concentration of the holes on P as well as N side should also become equal. In the context of energy band diagram, the Fermi level on N side should move down and the Fermi level on P side should move up, so that the Fermi levels on both the sides are equal. This will lead to a single Fermi level on both the sides indicating an equilibrium. In order to achieve all this, the electrons should diffuse from N to P side and the holes should diffuse from P to N side. In the beginning, the electrons and holes near the junction will start diffusing in to the opposite regions. The majority electrons on the N side are the ‘extra fifth’ and relatively free electrons associated with the dopant atom. Due to diffusion, the pentavalent dopant atom loses the electron and thus acquires a positive charge. Thus the dopant atoms near the junction will become positive ions which are heavy and thus immobile (Fig 6.21). When these electrons enter in P region, they recombine with holes. Similarly the diffusion of holes from P to N side creates an array of negative immobile near the junction on P side. After diffusion on N side, the holes recombine with the electrons. Thus the region near the junction becomes devoid of free charge carriers. Therefore this region is called as **depletion zone**. Typically the width of

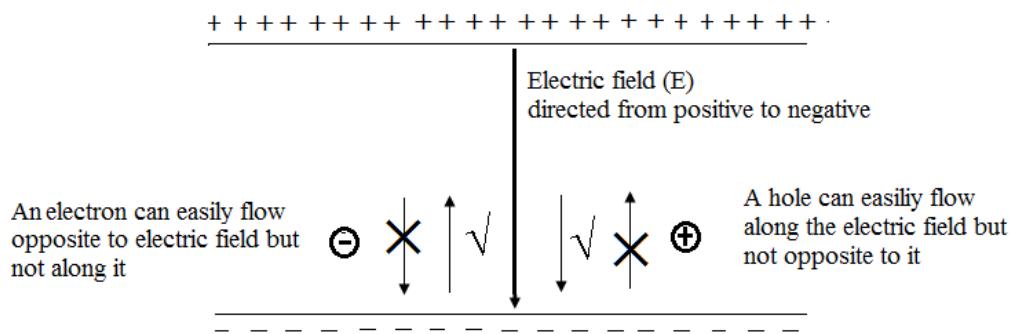


Figure (6.20): Response of electrons and holes to the electric field.

depletion zone is $1 \mu\text{m}$. As the depletion zone is devoid of free charge carriers, it has high resistance. The diffusion of electrons from N to P side (and holes from P to N side) gives rise to '**diffusion current (I_{diff})**'. This is also called as '**recombination current**' or '**majority current**'. The array of immobile ions in the depletion zone thus contains a space charge. The presence of positive immobile ions on N side and negative immobile ions on P side gives rise to a potential difference across the junction. The corresponding junction electric field is directed from positive (N) to negative (P). This junction-electric-field itself starts opposing the further diffusion of majority charge carriers. This is because electrons cannot flow along the electric field (and holes cannot move across the electric field) (refer Fig 6.20). The diffusion of electrons to the P side is opposed by the negative space charge near the junction. Similarly the diffusion of the holes in to

the N region is opposed by the positive space charge near the junction. As p.d. across the junction opposes the diffusion of majority charge carriers, this p.d. is also called as **barrier potential** (Fig 6.21). Thus, after the establishment of the p.d. across the junction the diffusion current (I_{diff} from N top P) decreases to a very small value. Now we may note that on P side, there are minority electrons (and on N side there are minority holes). Note that the junction electric field which opposes the diffusion of electrons from N to P region, itself acts as *emf* causing the drift of minority electrons from P to N region. Note that the electric field, which opposes the flow of electrons along its direction, itself favors their motion across it direction (Fig 6.20). In the similar manner, the junction electric field causes a drift of minority holes from N to P region. This results in to a small **drift current** (I_{drift}) directed from P to N. As the drift current is due to thermally generated minority charge carriers, it is also called as **minority current** or **thermal current**. Although the diffusion current is due to majority charge carriers, it is very small because it is opposed by the junction electric field. On the other hand, although the drift current is favored by the junction emf, it is small because it is due to the minority charge carriers.

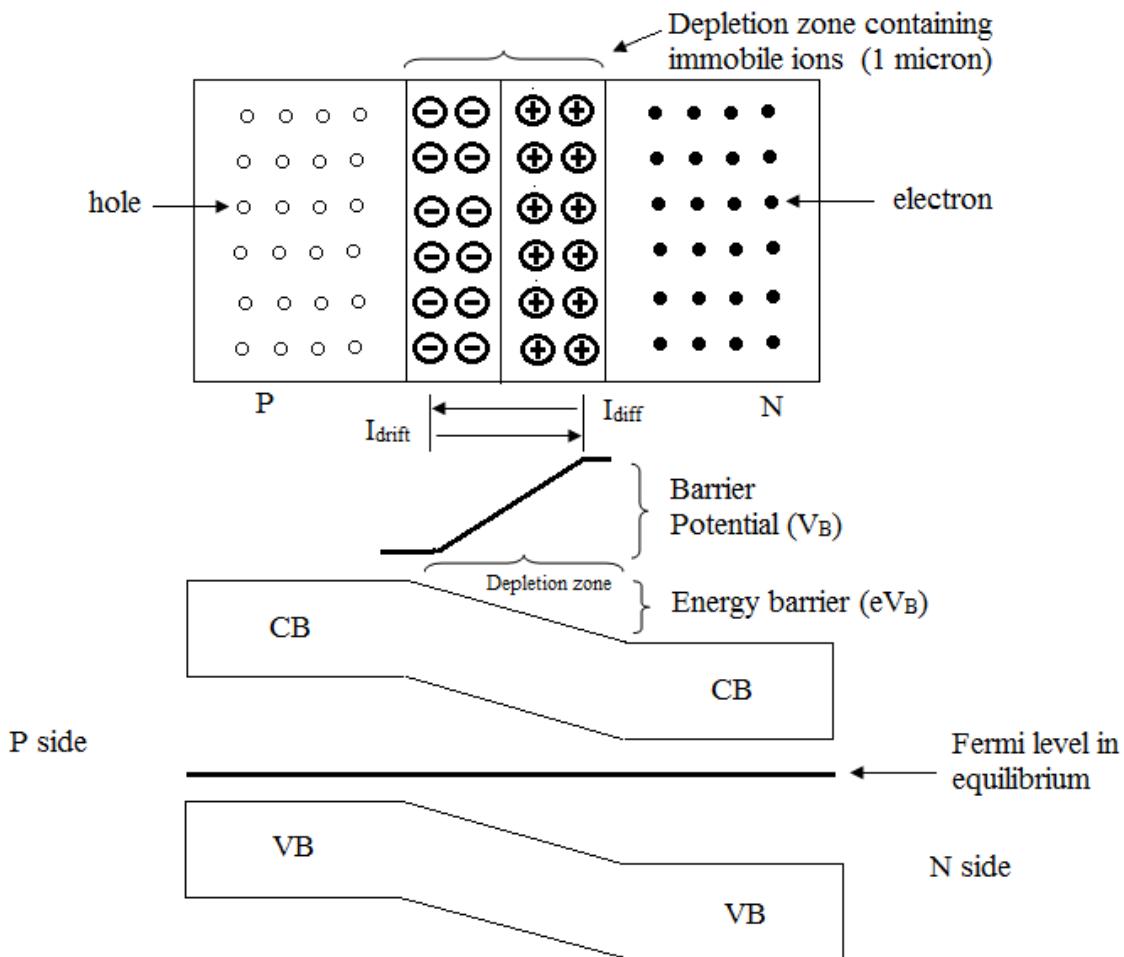


Figure (6.21): Energy band diagram of a PN junction diode, a few moments after its formation

Thus in unbiased (or open circuited) diode, the diffusion current and drift current are equal and opposite and thus the net current is zero (Fig 6.20). All these processes bring about a significant change in the energy band diagram. In the context of the Fermi level, the diffusion occurs due to difference in the positions of Fermi levels on N and P side. Thus the diffusion should continue till the Fermi levels on both the sides are equalized resulting in to a single Fermi level on both the sides. Consequently the diffusion should stop when the Fermi levels are equalized as the equilibrium is established. The diffusion of the electrons from N to P side brings down the Fermi level on N side (E_{FN}) and diffusion of holes from P to N side causes upward displacement of Fermi level on P side (E_{FP}). Note that the Fermi level is analogous to ‘center of gravity’ and it collectively represents the entire energy band diagram. Thus, when the Fermi level on N side moves down, it brings down the entire N-side energy band diagram. Similarly, when the Fermi level on P side moves up, it pulls up the entire P-side energy band diagram. This continues till the Fermi levels are equalized. Thus at equilibrium condition, though the Fermi levels on both the sides are at equal level, the corresponding energy band diagrams are not at the same level. The P-side energy bands are up and the N-side energy bands are down, and due to their relative positions, there occurs an energy barrier across the junction. As the barrier potential is V_B , the height of the energy barrier is eV_B . This energy barrier indicates that in the order to move from N to P side, the electrons should have an extra energy given by eV_B and as they don’t have it, their diffusion is stopped. The new energy band diagram also indicates that the electrons on P side are at higher energy levels as compared to N side.

PN junction diode under forward bias:

The energy band diagram of the unbiased PN junction diode shown in Fig (6.21) indicates the diode in equilibrium. Recall that in unbiased diode, the Fermi levels on both the sides are equalized. A diode can be biased in two ways; forward and reverse bias. In both cases, the equilibrium is disturbed and the diode responds in order to achieve equilibrium. In forward bias a battery (p.d. = V_F) is applied in such a way that positive terminal is connected to P region and negative is connected to N region. This process brings about significant changes in the diode and its energy band diagram. The negative terminal of the battery supplies electrons in N region. Secondly, the electric field of the external battery energizes the electrons on N side and pushes them towards the junction. Such electrons, which are pushed towards the junction, meet with the positive ions in the depletion region and neutralize them. Similarly the positive terminal of the battery increases the concentration of the holes in P region (or equivalently it pulls away the electrons from P region). The holes are pushed towards the junction. Such holes meet the negative ions in the depletion zone and neutralize them. Thus the immobile ions on both the sides are gradually neutralized. As a result, the depletion zone becomes successively narrower. Due to supply of electrons in N region and their energization, the Fermi level on N side shifts up, that is it moves towards the conduction band. On the other hand, the supply of holes in P region pulls the P side Fermi level down that is towards the valance band. Due to downward movement of P-side-Fermi-level and upward movement of N-side-Fermi-level, these two Fermi levels no longer remain at equal level and are separated by eV_F . Now the energy bands on N as well as P side adjust themselves to suit with the changes in Fermi levels. Thus the N side energy bands move up and P side energy bands move down. This reduces the energy barrier by eV_F . Thus now the decreased barrier is $eV_B - eV_F$. This reduction in the energy barrier can also be explained by recalling that the N side electrons are being energized by the external battery and depletion

region (containing space charge) is gradually narrowed. If the p.d. of the battery is further increased, the depletion zone vanishes and electrons easily move from N to P region. The reduction in the energy barrier enhances the current from N to P side. As discussed earlier, this is the diffusion current (or majority current or the recombination current). On the other hand the negligible drift current (or minority current or the thermal current) remains almost the same. Thus due to increased diffusion current, the current balance is disturbed and now the diode conducts with the net current given by $I_F = I_{diff} - I_{drift}$. This current is directed from N to P. The flow of the current can also be explained by considering that, the N-side-Fermi-level is up and the P-side-Fermi-level is down. This indicates a non-equilibrium situation and in order to achieve equilibrium, the electrons must flow from N to P side, so that the N-side-Fermi-level moves down and P-side-Fermi-level moves up. The equilibrium will be achieved only when these two Fermi levels are equalized. However this will never happen, as the electrons injected from N to P side will successively drift through P region and will leave it in order to meet the positive terminal of the external battery. Thus, as long as the external battery remains connected, the current will continue to flow. The substantial current of the diode in forward bias can also be understood by recalling that, depletion region, which offers high resistance, is diminished in this case.

PN junction diode in reverse bias:

When we connect the negative terminal of the external battery to P region and positive to N region, the diode is said to be reverse biased. The reverse biased PN junction diode behaves exactly opposite to forward biased PN junction diode. Recall that unbiased diode having equalized Fermi levels is in equilibrium condition. In reverse bias, this equilibrium is once again disturbed and the changes in the diode and its energy band diagram occur in order to bring about the equilibrium. As negative terminal of the battery is connected to P side, the supply and energization of the electrons occur on P side and hence the P-side-Fermi-level moves up. As the electrons in N side are pulled away and de-energized due to the positive terminal of the battery, the N-side-Fermi-level moves down. This creates a difference in these two Fermi levels, which is of the order of eV_R . Rest of the energy band diagram adjusts itself in order to be consistent with the changes in Fermi levels. Thus the N-side-energy-band-diagram moves down and P-side-energy-band-diagram moves up. This results in an increase in the energy barrier by eV_R and thus the net increased energy barrier becomes ($eV_B + eV_R$). This indicates that, now the electrons in N region face a greater energy barrier to enter in P region and as they don't have energy to do so, the diffusion current (or majority current or recombination current) falls to a negligible value. On the other hand, the drift current (or minority current or thermal current) remains almost the same. This is because, even though the external battery supports the flow of minority carriers from P to N region, the number of minority carriers (which depend only upon the temperature) remains the same. Thus now the new current equation is ($I_R = I_{drift} - I_{diff}$). As drift current is very less, the net current (I_R), which is directed from P to N is also feeble. Thus the PN junction diode does not conduct in reverse bias. Note that, upper position of P-side-Fermi-level and lower position of N-side-Fermi-level also indicates that, in order to achieve equilibrium, the electrons should move from P to N side. As these electrons drift through N region and meet the positive terminal of the battery, the equilibrium is never established and thus the reverse current continues to flow, as long as the battery is ON. In reverse bias the majority electrons in N region and majority holes in P region are pulled away by the external battery. As a result, the array of positive immobile

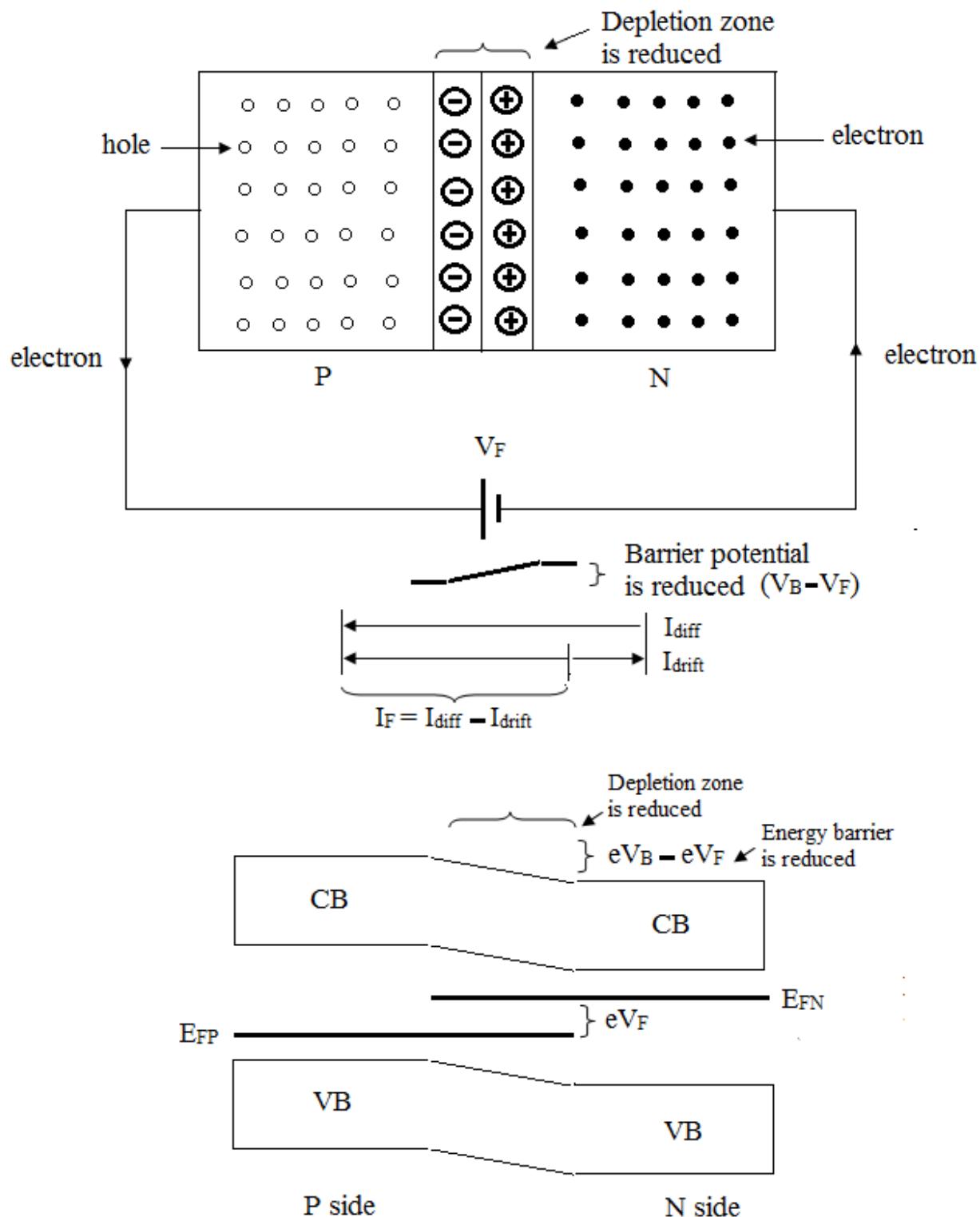


Figure (6.22): Energy band diagram of PN junction under forward bias

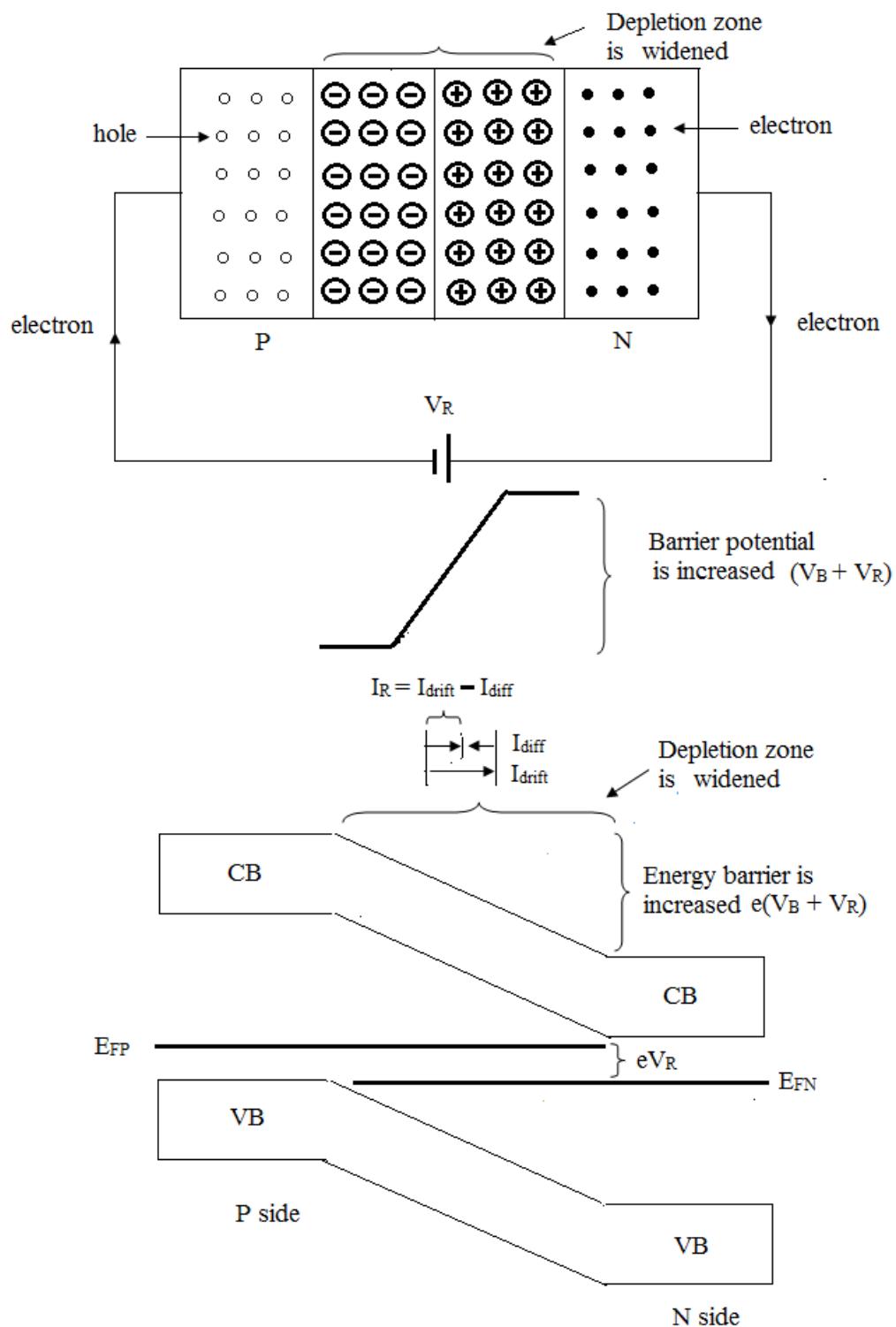


Figure (6.23): Energy band diagram of a reverse biased diode

ions on N side and negative immobile ions on P side expands and thus the depletion region which has high resistance is widened. Thus in reverse bias, the overall resistance of the diode is high. Widened depletion zone also indicates the increased value of barrier potential and energy barrier $\{e(V_B + V_R)\}$. It is to be noted that in forward bias, the external battery acts against barrier potential and in reverse bias it augments the barrier potential. The current in forward bias is dominated by majority charge carriers, while in reverse bias; it is dominated by minority charge carriers.

	<p>William Bradford Shockley (1910-1989): He was born in UK, but educated in California Institute of Technology (1932, graduation) and Massachusetts Institute of Technology (1936, Ph.D.). Afterwards he made his career in the famous Bell Telephone Laboratories and Shockley Semi-conductor Laboratory (where he invented transistor). Later on he became professor in Stanford University. He was also a visiting professor at Princeton University and California Institute of Technology. Transistor Physics was one of his several areas of interest. He made several inventions out of which, for 50 inventions he received patents and for one, transistor, the Nobel prize in 1956. He also published several articles and research papers.</p>
<p>John Bardeen (1908-1991): He was an American Physicist. His graduation and further career was in electrical engineering. However later he changed his interests to Physics and obtained graduate-degree and also Ph.D. in Physics in Princeton University. After a few years, he joined Bell Telephone Laboratories in 1945. Later on in 1951, he became Professor in University of Illinois. Out of several areas of his interest, one was electrical conduction in semiconductors and metals. In addition to his contribution in the invention of transistor (Nobel prize in 1956), he was one out of three Physicists who proposed BCS theory of superconductivity for which also, he was awarded Nobel prize in Physics in 1972</p>	
	<p>Walter Houser Brattain (1902-1987): He was an American Physicist. After graduation and post-graduation, he obtained Ph.D. from University of Minnesota in 1929. After 1929 he joined Bell Telephone Laboratories and there, he did research in several areas of Physics, especially on semiconductors, which resulted in the invention of transistor for which he shared Nobel prize in Physics in 1956.</p>

6.12 TRANSISTOR:

A Nobel Prize winning device which revolutionized the technology

Transistor was invented by William Shockley, John Bardeen and Walter Brattain in Bell Telephone Laboratories in 1947. It is invention of transistor which has made today's electronics compact, fast and versatile. Several areas including those from space exploration to atomic power and communications to computer and medicine have been benefitted due to invention of transistor. Nobel prize in Physics in 1956 was awarded to these inventors of the transistors. The

transistor that we discuss here is called BJT (Bipolar Junction Transistor). The word Bipolar indicates use of two charge carriers-electrons and holes while the word junction indicates use of two back to back PN junctions. There are other versions of transistor, namely FET, MOSFET and UJT and these operate by using only one kind of charge carrier. In electronics, transistors are used for two purposes; one as amplifier (current as well as voltage) and as a switch. Transistor plays a unique role in analog as well as digital electronics.

Structure of the transistor: There are two types of transistors; one NPN and another PNP. In NPN transistor, the majority charge carriers are electrons and in PNP they are holes. In NPN transistor, a lightly doped and thin P type of semiconductor is sandwiched between two N types of semiconductors, while in PNP transistor; a thin and lightly doped N type semiconductor is sandwiched between two P types of semiconductors. Thus transistor consists of three electrodes and two back to back PN junctions. The first electrode is called as emitter as it is the source of charge carriers (electrons in NPN and holes in PNP). The third electrode is called as collector, as it collects the charge carriers. The middle electrode is called as base. Emitter is heavily doped, as

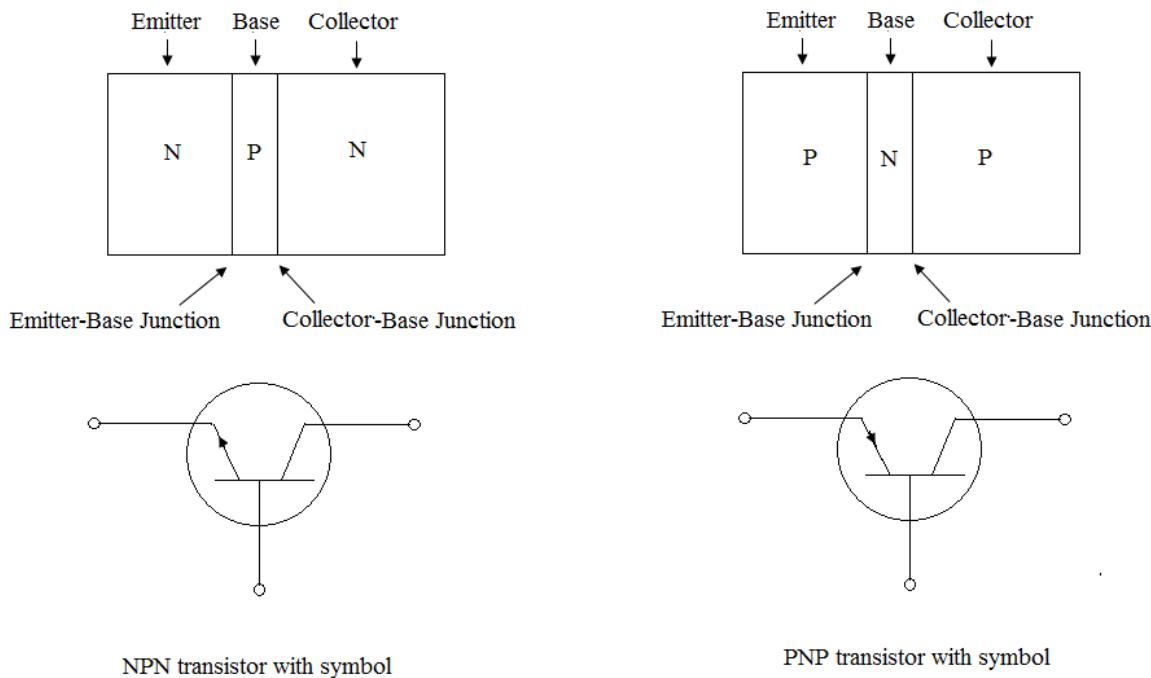


Figure (6.24) Transistor: Basic structure and symbols

it supplies the charge carriers. The base is extremely thin and lightly doped and collector is moderately doped. In NPN transistor, the base is a P electrode and in PNP transistor it is N. As a consequence, the base captures the charge carriers arriving from the emitter. However, as base is thin and lightly doped, it captures only a negligible number of charge carriers, which is nearly 2 %. Remaining 98 % charge carriers enter into the collector. Note that even though the transistor consists of two back to back connected PN junctions, the base of both these diodes is common. The ability of the transistor to amplify the signals depends upon how it is biased. The

emitter base junction is always forward biased and the base collector junction is always reverse biased. Due to forward biasing, the emitter-base has low resistance, while as collector-base junction is reverse biased, it has high resistance. Thus transistor transfers the resistance from low-resistance-input-emitter-base junction to high-resistance-output-base-collector-base-junction. Thus transistor is basically a transfer-resistor (transfer-resistor \Rightarrow transistor). Due to low resistance of input emitter-base junction, the voltage across it is low, while due to high resistance of the output collector-base junction, the voltage across it is high. This is the basic reason, why transistor works as an amplifier. As the base is lightly doped as compared to collector and emitter, the transistor can amplify the current also. As collector-base junction is reverse biased and as it offers high resistance, there occurs power dissipation across this junction due to which heat is generated. The collector is purposefully kept wide for dissipation of this heat. In both the transistors, the battery (V_{BE}) across emitter-base junction repels the charge carriers in the emitter towards the base. This constitutes, emitter current (I_E). Now as the base has opposite polarity, a few charge carriers (2 %) undergo recombination. These results in to base current (I_B). Rest of the charge carriers (98%) enter in the collector and afterwards they are attracted towards the battery (V_{CB}). Thus the fundamental current equation of the transistor is

$$I_E = I_B + I_C \quad \dots(6.37)$$

In order to differentiate between the symbols of NPN and PNP transistors, an arrow is indicated on the emitter electrode. For NPN transistor, the arrow points outwards, while in PNP, it points inwards. In both the cases the arrow indicates conventional current, which is opposite to the electron current. The transistor can be biased by keeping any electrode (emitter, base or collector) common to both the batteries. Accordingly there are three transistor circuit combinations, Common Emitter, Common Base and Common Collector.

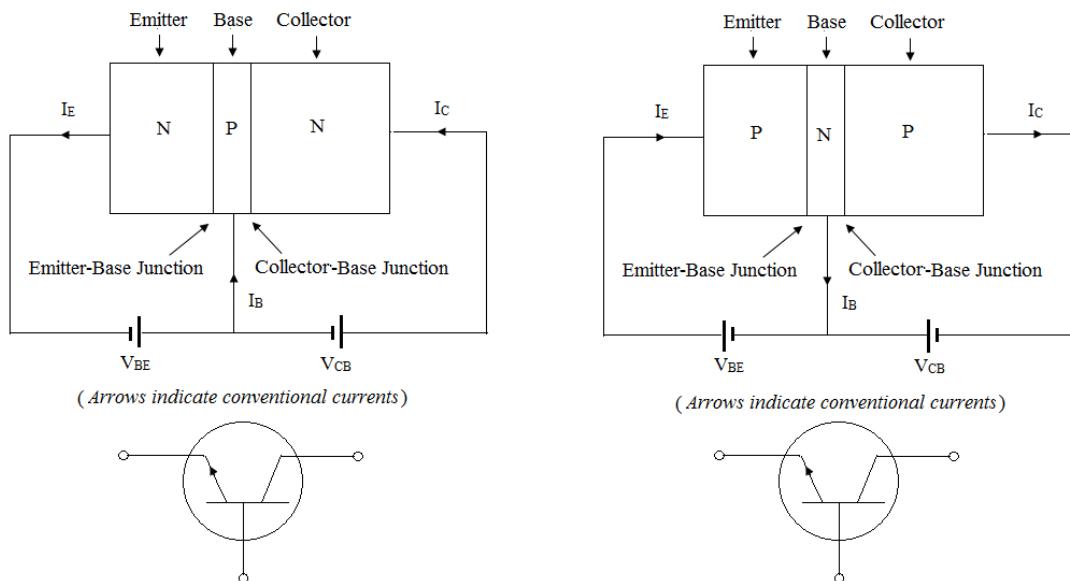


Figure (6.25): Biasing the transistor

Unbiased Transistor: Transistor consists of two back-to-back connected PN junctions with base electrode common. We know that a PN junction is characterized by depletion layers. Thus

transistor consists of two depletion layers, one across EB junction and another across BC junction. As base is thin and lightly doped as compared to emitter and collector, the sections of the depletion layers in emitter and collector are thin as compared to those in base. As base is already thin and then occupied by depletion layers, it gets encroached and becomes still thin. However, base, although thin and lightly doped, plays an important role in transistor action.

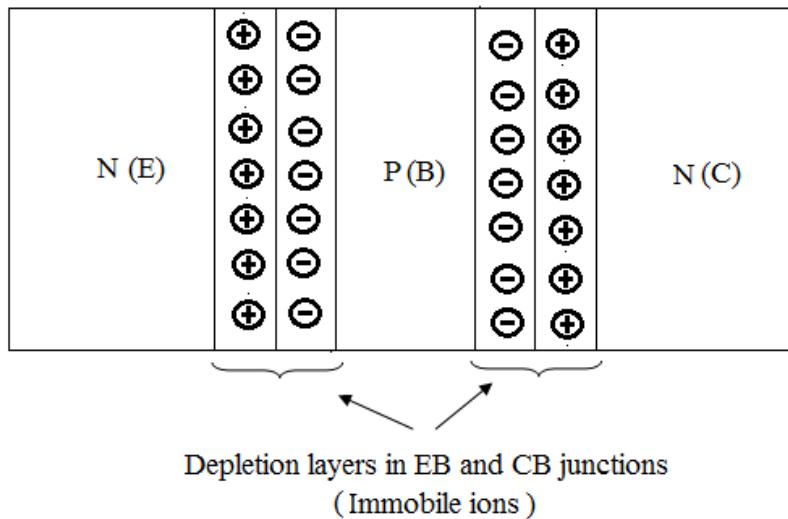


Figure (6.26): Depletion zone is transistor

Energy Band Diagram of Unbiased NPN Transistor

Energy band diagram of an unbiased NPN transistor at the instant of its formation ($t = 0$)

Consider a transistor just at the moment of its formation (refer Fig (6.27)). As such, transistor is not made by joining emitter, base and collector. A single chip of a pure semiconductor is doped with suitable pentavalent and trivalent impurities at appropriate doing levels. Consider NPN transistor. Here emitter and collector are of N type. Thus due to high concentration of free electrons in conduction band, the Fermi levels of both these electrodes are near the conduction band. This indicates the high concentration of energetic electrons (in conduction band) in these electrodes. Base is made up of P type semiconductor. Due to higher concentration of holes in the valance band, the Fermi level of base electrode is near the valance band. This also indicates the absence of energetic and free electrons (in the conduction band). The presence of these Fermi levels at different positions indicates the non-equilibrium condition and this indicates that a set of processes must take place to bring all Fermi levels at the same level

Energy band diagram of an unbiased NPN transistor a few moments after its formation

The energy band diagram of a transistor can be readily understood if we consider that a transistor is a combination of two back to back PN junction diodes. Fig (6.28) indicates that the Fermi levels of N type emitter and N type collector are at higher level (close to conduction band)

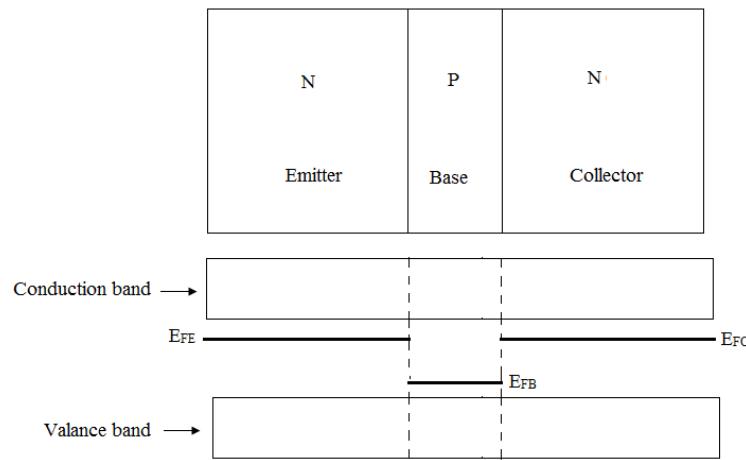


Figure (6.27): Energy band diagram of NPN transistor at the instant of its formation ($t = 0$)

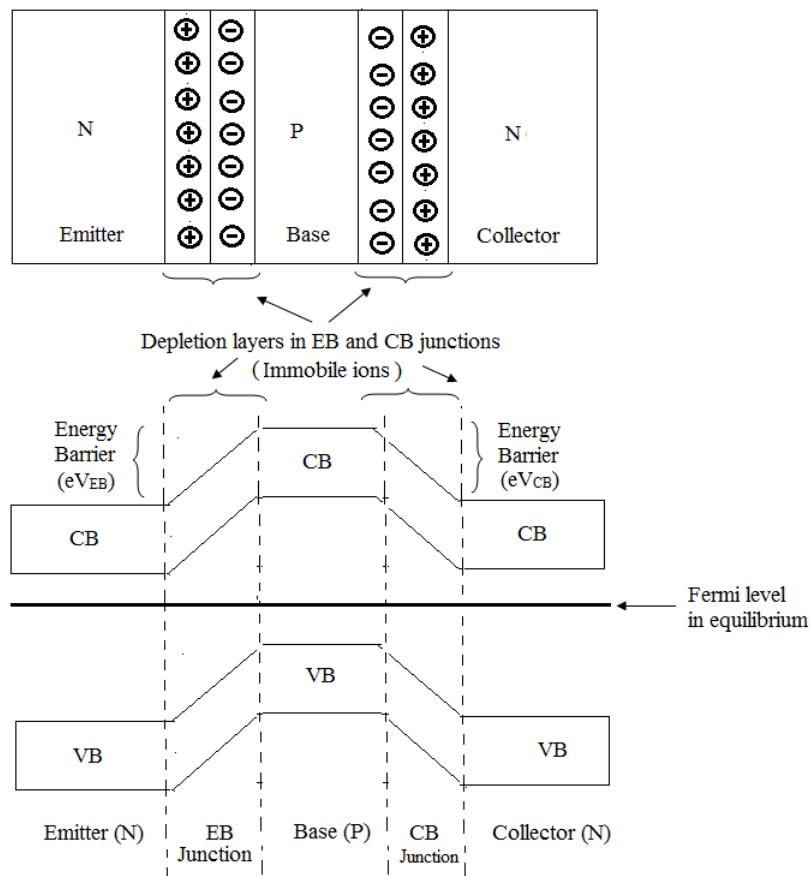


Figure (6.28): Energy band diagram of unbiased NPN transistor, a few moments of its formation

and that of P type base is at lower level (close to valance band). These positions signify the high concentration of electrons in emitter and collector and high concentration of holes in base. This is a non-equilibrium situation. The equilibrium requires that the Fermi levels in all the regions be equalized. For this to happen, the Fermi levels of emitter and collector need to move down and the Fermi level of the base needs to move up. Consequently the electrons in emitter (near the junction) and the collector (near the junction) diffuse in the base. The atoms near the junction from which these electrons leave become positive ions. The diffusion of electrons from emitter and collector to base and the diffusion of holes from base to emitter and collector results in to formation of negative immobile ions in the base near the junctions. As discussed earlier in the case of diode, this results in the formation of voltage barriers across the emitter-base and collector-base junctions. These barriers oppose the further diffusion of electrons from emitter and collector to the base and holes from base to emitter and collector. In energy band diagrams this indicates an equilibrium situation in which the Fermi levels in all the regions are equalized resulting in to a single Fermi level of the entire transistor. As Fermi level represents the energy status of a band, the lowering of Fermi levels in emitter and collector brings down the band structures of these electrodes. On the contrary, the upward movement of the Fermi level of the base pulls the band in upward direction. As a result, there occurs bending of the bands and the energy band diagram acquires a structure as shown in the Fig(6.28). The bending of bands (downward movement of bands of emitter and collector and upward movement of the bands of the base) results in to the formation of the energy barriers with heights eV_{EB} in emitter- base junction and eV_{CB} in collector-base junction. This also indicates that the electrons in the emitter and collector are at lower energy levels than those in base and they need to climb the energy barriers of eV_{BE} and eV_{CB} for entering emitter or collector into the base. As these electrons are devoid of the required energies, the current ceases. This can also be understood by recalling that the Fermi levels in all the regions are equalized. It can be seen that even in equilibrium condition, where all Fermi levels are equalized, the Fermi level is close to the conduction bands of emitter and collector and close to the valance band of base. The entire discussion leads to a conclusion that an unbiased transistor does not conduct.

Biassing NPN transistor: As discussed earlier, transistor consists of two back-to-back connected PN junction diodes. We also know that a diode can be biassed in two ways; forward and reverse. Thus, for biassing the transistor there can be four permutations and combinations. However, for transistor to work as an amplifier, the emitter-base junction needs to be forward biassed and the collector-base junction needs to be reverse biassed.

As transistor consists of two back-to-back connected diodes, the energy band diagram of a biassed transistor can be easily understood by recollecting the energy band diagrams of the forward and reverse biassed PN junction diodes. We know that when PN junction diode is forward biassed, the energy barrier decreases while if it is reverse biassed, the energy barrier increases. Refer Fig (6.29). As emitter-base junction is forward biassed, the energy barrier across the junction decreases by an amount ($eV_{BE} - eV_{BEext}$). On the contrary, as the collector-base junction is reverse biassed, the energy barrier increases by an amount ($eV_{CB} + eV_{CBext}$). Further, as the negative terminal of battery (V_{BE}) is connected to N type emitter, the electrons in the emitter are energized and their concentration also increases. As a result, the Fermi level of the emitter is raised by eV_{BEext} . This Fermi level, while rising upward, also pulls up the energy band diagram of the emitter. This also explains the reduction of the energy barrier across emitter-base junction. The Fermi level and the entire energy band diagram of the base slightly moves down

due to the effect of the positive terminal of the battery applied to the base. Collector-base junction is reverse biased and thus the N type collector is applied with the positive terminal of the external battery. This depletes the concentration as well as the energy of the electrons in the collector. As a result

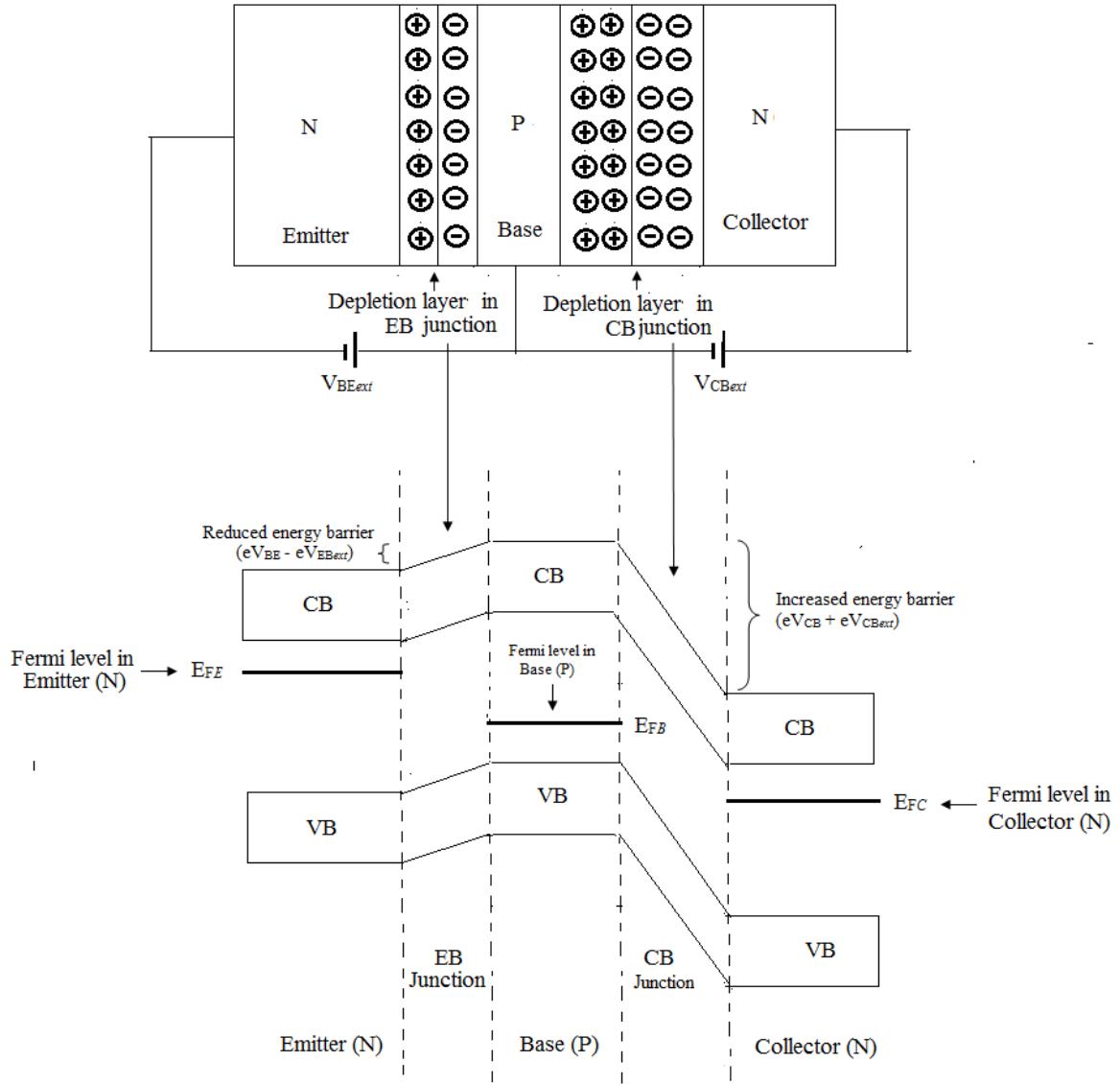


Figure (6.29): Energy band diagram of a biased NPN transistor

the Fermi level of the collector side moves down by an amount eV_{CBext} . This Fermi level, while moving down, pulls down the energy band diagram of the collector side in downward direction. This also explains why energy barrier across the collector-base junction is increased. Note the relative positions of the Fermi levels of emitter (E_{FE}), base (E_{FB}) and collector (E_{FC}). E_{FE} is at higher position, E_{FB} is intermediate position and E_{FC} is at the lower position. This indicates that in order to bring about the equilibrium(that is the equalization of the Fermi levels of the emitter,

base and collector side), the electrons must flow from emitter to the base and then base to the collector.

The decrease in the energy barrier of the emitter junction, raised position of the Fermi level of emitter side and the forward biasing of emitter-base junction collectively indicate that, emitter should inject electrons in the base region. This emitter to base current is termed as I_E . The electrons injected from emitter enter in the base and, as base is P type, they combine with the holes. But the base is thin and lightly doped, therefore only few electrons (roughly 2%) recombine with the holes and remaining 98% electrons move towards the collector. The 2 % electrons recombined with the holes are attracted by the positive terminal of the battery connected to the base. This negligible current coming out of the base is called as base current (I_B). The 98 % electrons entering in to the collector now face a downward journey across the base-collector energy hill. (Note that electrons cannot climb up the energy barrier, but their downward journey on the energy barrier is quite easy). The 98 % electrons entered in the collector are pulled away by the positive terminal of the external battery (V_{CBext}) and they constitute the collector current (I_C). The journey of the electrons from the emitter to base and then from base to collector can also be understood by the relative positions of the Fermi levels on emitter, base and collector side. The electrons leaving the collector, add up in the electrons coming out of the base, and these combined electrons reenter in to the emitter. Thus the basic current equation of transistor is

$$I_E = I_B + I_C$$

As there is a continuous injection of the electrons in to the emitter and continuous ejection of the electrons from the collector, the Fermi levels of these electrodes are never equalized and thus due to their relative positions, the current continues to flow as long as the circuit is ON.

The energy band diagram of PNP transistor can be treated as a mirror image of the energy band diagram of NPN transistor.

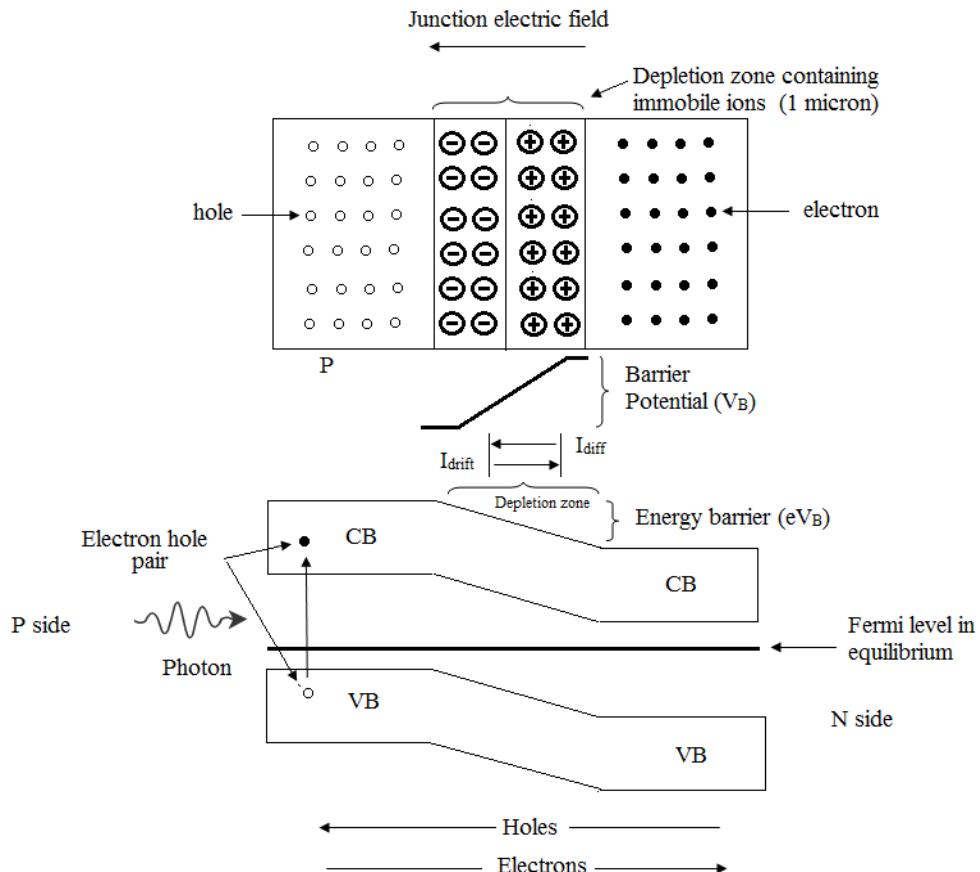
 Gerald Pearson (left) with his colleagues in Bell lab	Gerald Pearson (1905-1987): He was an American Physicist who is credited for the invention of solar cell. He obtained degree from Stanford University and afterwards joined Bell laboratories as a Physicist. There he earned several patents while working on PN junctions and transistor. There was no photovoltaic industry before 1954, however, in 1954, along with his colleagues Daryl Chapin and Calvin Fuller,Pearson invented the first silicon solar cell. The first solar cell had an efficiency of only 6%, however due to intense research thereafter, the efficiency increased to 10 %. After 1960, Pearson joined Stanford University as a Professor of Physics
--	--

6.13 SOLAR CELL:

A PN junction diode can convert sunlight in to electricity

We know that a PN junction diode has two electrodes, P and N, and a barrier potential across the

junction. This barrier potential corresponds to a junction electric field which is directed from N to P. When a PN junction is exposed to light (Fig 6.30), the photons excite the electrons in the valance band in to conduction band. These photonically generated electrons and holes move in opposite direction due to the action of the junction electric field. (We know that an electron flows opposite to electric field while a hole moves in the direction of electric field). Thus all photonically generated electrons in P region are swept in N region and all photonically generated holes are swept from N region to P region (Fig 6.30 and 6.31). The electrons entered in N



(6.30): Solar cell and its energy band diagram

Figure

region continue to flow towards the surface and they accumulate in the metal contact provided on the surface (Fig 6.31). Similarly holes entered in to P region continue to flow towards the surface and accumulate in the metal contact on the surface. Thus the N side metal contact acquires negative potential and P side metal contact acquires positive potential and consequently a potential difference is created across the diode. If the diode is connected to a load then this PD drives a current in the circuit. Thus we get electrical power. (Fig) Such diodes which are capable of converting light in to electricity are called as solar cells. The phenomenon of converting the light in to voltage is called as photovoltaic effect and thus solar cells are also called as photovoltaic cells. Sometimes solar cells are also called as solar batteries as they give

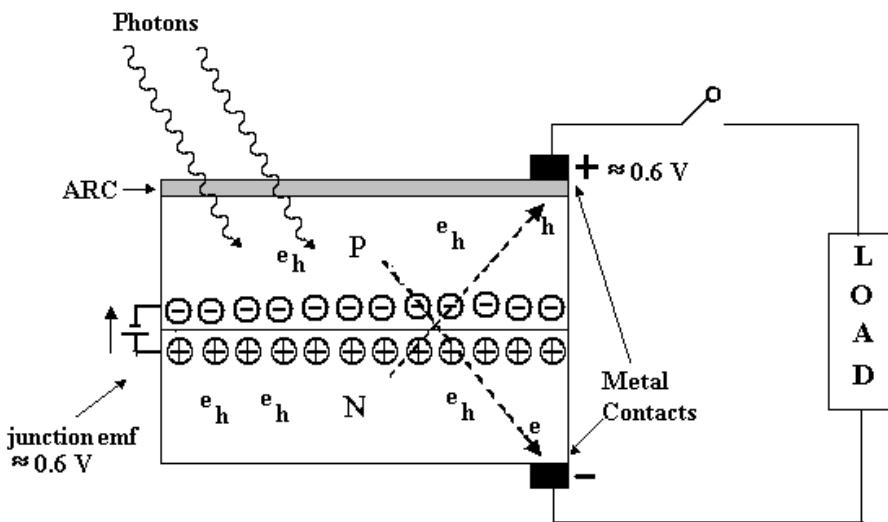


Figure (6.31): Solar cell and its working

electrical power. Thus solar cell generates the electrical power in four steps

- i. Generation of electrons and holes due to light
- ii. Separation of these electrons and holes due to junction- electric-field
- iii. Their accumulation across the metal contacts and thus the generation of emf
- iv. Flow of current due to this emf, when solar cell is connected across a load

For the appropriate conversion of light in to electricity, the area of the diode exposed to the light should be as large as possible. Further, to prevent the recombination of electron and holes before their separation, the maximum number of electron hole pairs must be generated near the junction (depletion region). For this purpose, the P electrode which faces the incident light is made quite thin. Thus solar cell is a modified PN junction diode. The efficiency of the solar cell, when it was just invented in 1954, was low; it was roughly 6%. Due to an intense research thereafter, at present, the efficiency of solar cells is in the range of 10-30%. As such, the output of a single PN junction solar cell is small; the voltage is approximately 0.5 and the current is in the range of few microamperes. For enhancing the electrical output of solar cell, several solar cells are cascaded in series or parallel, as per the requirement. Such combinations of solar cells are called as solar photovoltaic panels (or modules or arrays).

I-V Characteristics of Solar cell: A typical circuit that can be used to study I-V characteristics of solar cell is shown in Fig (6.32). When the load is not connected (or connected, but very high), the current in the circuit is zero. Consequently the voltage across the cell is maximum. This is called as *open circuit condition* and the corresponding voltage is called as *open circuit voltage (V_{OC})*. On the contrary, if the load resistance is reduced to zero, maximum current flows through the circuit, but then the voltage drops to zero. This is called as *short circuit condition* and the corresponding current is called as *short circuit current(I_{SC})*. Open circuit condition corresponds to infinite load and short circuit condition corresponds to zero load. Therefore both

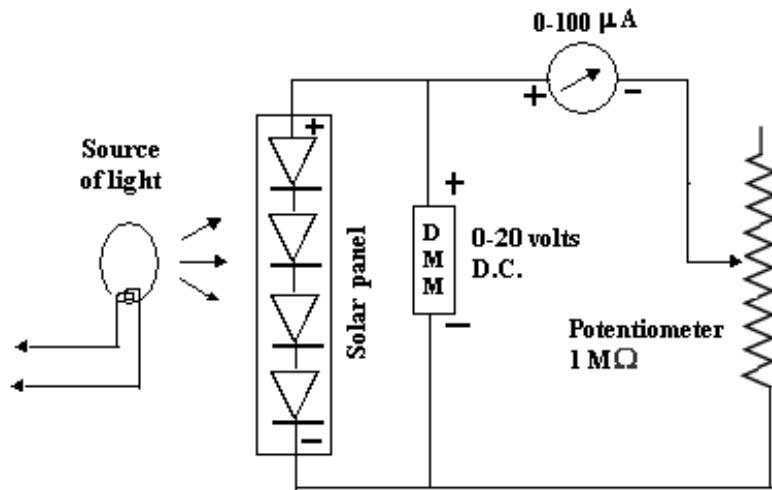


Figure (6.32): A typical circuit that can be used to study I-V characteristics of solar cell

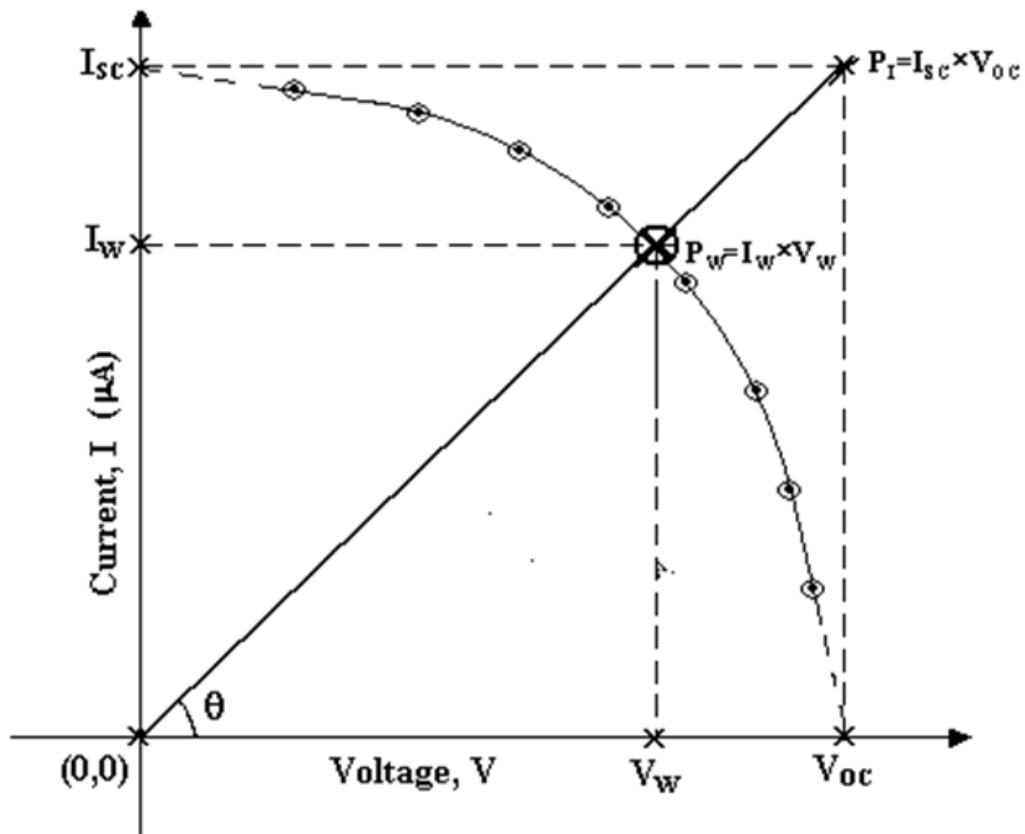


Figure (6.33): I-V characteristics of a solar cell

these conditions cannot be achieved in practice. In practice, while recording the currents and voltages from solar cell, when load is kept maximum, the corresponding voltage is maximum and very close to V_{OC} while when the load is minimized, the corresponding current is maximum and very close to I_{SC} (refer Fig 6.33). The actual V_{OC} and I_{SC} can then be obtained by extrapolating the characteristics on voltage and current axes. (refer Fig 6.33). As both I_{SC} and V_{OC} are maximum amongst all possible currents and voltages from solar cell, the product $I_{SC} \times V_{OC}$ represents maximum possible power from the solar cell. However this power is not achievable due to two reasons, one, ideal open circuit condition and ideal short circuit condition cannot be achieved in practice, and second, both these conditions cannot be obtained simultaneously. Thus this power is called as *ideal power*. We have

$$\text{Ideal power} = P_I = I_{SC} \times V_{OC} \quad \dots (6.38)$$

In the Fig. (6.33), ideal power (P_I) signifies the area under *ideal power rectangle*. Also note that the *ideal power point* having coordinates (V_{OC}, I_{SC}) , does not lay on the actual I-V characteristics. This also signifies that ideal power is unachievable. All the powers laying on the I-V curve are achievable i.e. real. Amongst these, the maximum (and at the same time real) power can only be obtained on a point somewhere in the middle portion of the graph. The powers too close to current and voltage axes are also real but not workable, as there, either current or voltage is small. The region, where the current and voltage are simultaneously optimum is expected to occur somewhere in the middle of the characteristics. If the scales on X axis and Y axis are equivalent, then the characteristics is symmetric. In such case, the optimum point can be obtained simply by drawing a line inclined at 45° from origin. The point where it meets the curve is expected to give optimum point. However, the scales are not uniform. The optimum point deviates from the middle. In such case, the optimum point can be obtained by using following procedure

- i Draw a perpendicular on Y (current) axis at I_{SC} . Also draw a perpendicular to X (voltage) axis at V_{OC} .
- ii These perpendiculars meet at a point whose coordinates are (V_{OC}, I_{SC}) . As discussed earlier, this is an ideal power point and the area under ideal power rectangle formed by above perpendiculars and X, Y axes signifies the ideal power.
- iii If this ideal power point is joined to the origin by drawing a line, then no matter what the scale is or no matter whether the graph is perfectly symmetric or not, this line always intersects the curve at optimum power point. As at this point, the current and voltage are simultaneously optimum. As this power is real and optimum, it is also called as *workable power* (P_W). The corresponding *workable voltage* (V_W) and *workable current* (I_W) can be obtained by drawing perpendiculars from workable power point on voltage and current axes respectively. Thus we have

$$\text{Workable power} = P_W = I_W \times V_W \quad \dots (6.39)$$

- iv The workable power corresponds to the area under workable power rectangle (as shown in the Fig 6.33)
- v As we can see, the workable power rectangle is smaller than ideal power rectangle (and it

should always be). However, some fraction of ideal power rectangle is ‘filled’ by workable power rectangle. Greater the ‘filling’, more close is the workable point to the ideal power point. This can be described by introducing a physical quantity called as *fill factor*. We have

$$\text{fill factor} = f = \frac{P_W}{P_I} \times 100 \% \quad \dots(6.40)$$

- vi Thus if the fill factor is 60%, then it means that the workable point is 60% close to (or only 40% away from) ideal power point. Thus fill factor is the measure of the quality of the solar cell-load circuit.
- vii The I-V characteristic of the solar cell is obtained by varying the load resistance. The power output of the solar cell is low for minimum or maximum load, as consequently for these loads, either voltage or current is minimum. Thus the load should be optimum. This load can be easily calculated by taking the ratio of workable voltage and workable current. We have

$$\text{Workable load} = R_W = \frac{V_W}{I_W} \quad \dots(6.41)$$

- viii The workable load signifies the resistance which any application (to be driven by the given solar cell) should have if it has to extract maximum power from the solar cell
- ix It may be noted that, though fill factor compares two powers; the workable and ideal, it differs from the efficiency of the solar cell. Fill factor compares the powers, both of which are output and both of them are electrical. In efficiency, we compare the output electrical power with the input optical power. We have

$$\text{efficiency} = \varepsilon = \frac{P_{\text{output}}}{P_{\text{input}}} = \frac{P_{\text{workable}}}{P_{\text{optical}}}$$

Merits and Demerits of Solar cells: Though nanotechnology promises the solar cells with 90% efficiency in the future, the efficiency of the present solar cells at commercial level is only around 10%. Solar electricity has its own merits and demerits, some of which are described below

Merits:

- i. Solar cell converts sunlight into electricity. Sunlight is abundantly available and it is inexhaustible. The other forms of fuels such as, coal, uranium, natural gas, diesel, petrol are exhaustible and they are depleting rapidly.
- ii. Solar cell does not create any pollution. There is no global warming. In all other methods of generating electricity such as coal, uranium, natural gas, diesel, petrol there is either chemical or radioactive pollution. This also results in global warming
- iii. Solar electricity is risk free electricity. It is safe. There is no fear of accidents. The accident in thermal or nuclear reactors or hydraulic power can lead to a disaster.

- iv. Especially in case of hydraulic power, dams are necessary. This leads to displacement of people and their rescue. Such problems do not occur in solar electricity.
- v. In all other options of electricity generation, say thermal, nuclear or hydraulic, huge and skilled manpower is required which is not necessary in solar power plant. People in villages can also be trained to operate and maintain the solar power plants
- vi. Solar electricity can be generated in centralized as well as decentralized manner and in the latter case; the solar panels can be fitted on houses, street lights or agricultural pumps. This reduces the transmission losses to a great extent. In other options, the electricity is generated in a centralized manner and then distributed at required places using cables. Almost 40% electricity is lost during such transmission (I^2R losses).
- vii. The process of conversion of sunlight in to electricity occurs instantly. There are no stages involved , which is the case in other options
- viii. Low maintenance

Despite the attractive merits as mentioned above, solar electricity has not yet become a unique and completely reliable option of electricity generation. This is because, at present, solar electricity has a few demerits which are mentioned below.

Demerits:

- i. The sunlight follows day night cycle. The availability of sunlight is also affected due to change in seasons, change in climate and cloudy weather. Thus solar electricity cannot be generated in a 24×7 manner
- ii. The electricity generated by solar cells in the day period can be stored and then used in the night. But the storage methods are costly.
- iii. The efficiency of solar cells is limited (around 10% at present at commercial level). This results in to increase in the overall cost. Thus the solar electricity is not cheap electricity.
- iv. Solar electricity is generally preferred for low power applications. It is not used when heavy power is required. As such, in most of the cases, solar electricity is a weak electricity.
- v. In order to enhance the electrical output in spite of low efficiency, the area of solar panels needs to be as large as possible. This sometimes results in to occupation of land
- vi. Solar cells generate DC electricity. In many cases, especially for effective transmission, AC electricity is required. Therefore conversion of electricity from DC to AC is required. This requires additional facilities.

Applications of the solar-photovoltaic cells: The applications of solar cells have three major aspects.

- i. Though it is not possible to completely rely on solar electricity (due to above-mentioned demerits), it can still be used to reduce the load on conventional electricity.

- ii. In some applications, especially satellites and remote areas such as villages and deserts, high altitude places in Himalayas, Antarctica etc., the conventional electricity cannot be used due to transportation problems. In such places, solar electricity is inevitable. Indeed, the first applications of solar cells involved satellites only. The demerits of solar cells such as low efficiency and high cost are secondary in case of satellites. Reduction of weight is a primary factor and solar cells are preferred due to their light weight.
- iii. Some low power devices such as toys, calculators, emergency lights, LEDs, chargers solar cells can be used quite effectively.

A few present and future applications of solar cells are mentioned below

- i. Supply of electricity in remote areas including villages and deserts, high altitude places in Himalayas, Antarctica, where conventional electricity cannot be transmitted
- ii. Satellites are the unique application of solar electricity
- iii. Low power devices such as calculators, toys, LEDs, chargers etc.
- iv. Street lights, marine lights, lights in airports
- v. Street signals, railway signals etc.
- vi. Agricultural pumps
- vii. Solar villages
- viii. Solar photovoltaic power plants
- ix. Instant supply of electricity in disasters such as earthquakes and floods
- x. TV transmitters
- xi. Powering medical equipment in remote areas
- xii. Solar bicycles, solar cars, solar boats, solar aircrafts etc.
- xiii. Alarms at remote places

A heavy research is going on solar cells and their applications to the mankind.



***World's largest Solar Photovoltaic Power Plant is situated in Madhya Pradesh, India.
(750 MW, 1,500 hectare and Rs. 45 billion). .
Semiconductor Physics is the basis of world of electronics***



Silicon diode



Germanium diode



Zener diode



Thermistor



LED



Photodiode



LDR



Solar cell



Bipolar junction
transistor (BJT)



Field Effect
Transistor (FET)



Silicon Controlled
Rectifier (SCR)



(Metal-Oxide
Semiconductor Field-Effect
Transistor) MOSFET

MULTIPLE CHOICE QUESTIONS

Q 1 Semiconductor devices have following advantages over vacuum tube devices

- a They consume less power
- b They are compact
- c They are fast
- d All of a, b and c

Q2 The atomic number of Germanium is

- a 25
- b 31
- c 30
- d None of a, b and c

Q 3 Spin of the electron is

- a $+\frac{1}{2}$
- b $-\frac{1}{2}$
- c $-\frac{1}{2}$ or $+\frac{1}{2}$
- d $-\frac{1}{2}$ and $+\frac{1}{2}$

Q 4 Which of the following is not a semiconductor

- a Germanium
- b Silicon
- c Gallium Arsenide
- d None of a, b and c

Q 5 Which of the following is not a trivalent impurity?

- a Aluminium
- b Boron
- c Indium
- d None of a, b and c

Q 6 Which of the following is not a pentavalent impurity?

- a Phosphorus (P)
- b Arsenic (As)
- c Gallium (Ga)
- d Bismuth (Bi)

Q 7 As the atoms of a crystalline solid come together, their energy levels split and become bands. This is because of

- a Pauli's exclusion principle
- b Overlapping of the wavefunctions of the electrons
- c Interaction between the electrons
- d All of a, b and c

Q 8 Valance band is the

- a Highest occupied band
- b Lowest unoccupied band
- c Highest unoccupied band
- d Lowest occupied band

- Q 9** Conduction band is the
- a Highest occupied band b Lowest unoccupied band
c Highest unoccupied band d Lowest occupied band
- Q 10** In the energy band diagram of the diamond, the valance band is
- a Occupied 2s-2p hybrid band b Unoccupied 2s-2p hybrid band
c Unoccupied 3s-3p hybrid band d Occupied 3s-3p hybrid band
- Q11** The energy gap of Germanium is 0.7 eV. It can absorb a radiation having wavelength
- a 18903 \AA° b 17760 \AA°
c 5500 \AA° d 6328 \AA°
- Q 12** The energy gap of Silicon is 1.1 eV. When Silicon diode is forward biased, the electrons in the conduction band recombine with the holes in the valance band. Due to this, following radiation is emitted
- a Light b Ultraviolet
c Infrared d Microwave
- Q13** Silicon doped with Aluminium behaves as a
- a P type of semiconductor b N type semiconductor
c Intrinsic semiconductor d None of a, b and c
- Q 14** Silicon doped with Phosphorus behaves as a
- a P type of semiconductor b N type semiconductor
c Intrinsic semiconductor d None of a, b and c
- Q 15** In semiconductor, donor levels are the energy levels which appear
- a Near the valance band and created due to doping of trivalent impurity b Near the conduction band and created due to doping of pentavalent impurity
c Near the conduction band and created due to doping of trivalent impurity d Near the valance band and created due to doping of pentavalent impurity

Q 16 In semiconductor, acceptor levels are the energy levels which appear

- a Near the valance band created due to doping of pentavalent impurity
- b Near the conduction band created due to doping of pentavalent impurity
- c Near the valance band created due to doping of trivalent impurity
- d Near the conduction band created due to doping of trivalent impurity

Q 17 At 0 K, the Fermi level is a

- a Highest occupied energy level at the top of valance band
- b Lowest occupied energy level at the bottom of the conduction band
- c Highest unoccupied energy level at the top of the valance band
- d Lowest unoccupied energy level at the bottom of the conduction band

Q 18 At any finite temperature, the Fermi level in semiconductor is a level having probability of occupancy

- a Less than 50 % and it is near the valance band
- b Equal to 50 % and it is at the center of the energy gap
- c Greater than 50% and it is near the conduction band
- d Equal to 100 % and it is at the top of valance band

Q19 The Ohm's law in terms of J , E and ρ is given by

- a $E = \frac{J}{\rho}$
- b $E = \frac{\rho}{J}$
- c $E = J \times \rho$
- d $\rho = J \times E$

Q 20 The Fermi Dirac probability function is given by

- a $P(E) = \frac{1}{1 - e^{\frac{(E-E_F)}{KT}}}$
- b $P(E) = \frac{1}{1 - e^{\frac{(E+E_F)}{KT}}}$
- c $P(E) = \frac{1}{1 + e^{\frac{(E-E_F)}{KT}}}$
- d $P(E) = \frac{1}{1 + e^{\frac{(E+E_F)}{KT}}}$

Q 21 Resistivity is given by

- a $\rho = \frac{R}{l/A}$
- b $\rho = \frac{R}{A/l}$
- c $\rho = \frac{R \times l}{A}$
- d $\rho = R \times l \times A$

Q 22 Mobility is defined as

- a Drift velocity per unit potential difference
- b Drift velocity per unit electric field
- c Current density per unit electric field
- d Current per unit potential difference

Q 23 According to Hall effect

- a If a current carrying specimen is kept in the longitudinal electric field then a potential difference is induced in a direction perpendicular to the current and electric field
- b If a current carrying specimen is kept in the longitudinal magnetic field then a potential difference is induced in a direction perpendicular to the current and magnetic field
- c If a current carrying specimen is kept in the transverse electric field then a potential difference is induced in a direction perpendicular to the current and electric field.
- d If a current carrying specimen is kept in the transverse magnetic field then a potential difference is induced in a direction perpendicular to the current and magnetic field

Q 24 The Hall voltage is given by

- a $V_H = \frac{1}{nq} IB \frac{d}{A}$
- b $V_H = R_H IB \frac{d}{A}$
- c $V_H = \frac{1}{nq} JB d$
- d All of a, b and c

Q 25 The resistivity of the semiconductor

- a Decreases exponentially with the temperature
- b Increases with temperature
- c Remains constant with temperature
- d Decreases linearly with the temperature

Q 26 The energies of the electrons in the metals are distributed according to

- a Fermi-Dirac distribution
- b Bose-Einstein distribution
- c Maxwell-Boltzmann distribution
- d Random distribution

Q 27 The energies of the electrons in the semiconductors are distributed according to

- a Fermi Dirac distribution
- b Bose-Einstein distribution
- c Maxwell-Boltzmann distribution
- d Random distribution

Q 28 The conductivity of the conductor is given by

- | | |
|--------------------|----------------------|
| a $\sigma = ne\mu$ | b $\sigma = neE$ |
| c $\sigma = neJ$ | d None of a, b and c |

Q 29 In Hall effect, the current is along X axis and if the magnetic field is along Z axis then, the Hall voltage is along

- | | |
|----------|-------------------------------------|
| a Z axis | b Y axis |
| c X axis | d At 45° w.r.t. X and Z axis |

Q 30 The Fermi level of N type of semiconductor is

- | | |
|------------------------------|---------------------------------|
| a Near the conduction band | b Near the valance band |
| c Within the conduction band | d In the middle of the band gap |

Q 31 The Fermi level of P type of semiconductor is

- | | |
|------------------------------|---------------------------------|
| a Near the conduction band | b Near the valance band |
| c Within the conduction band | d In the middle of the band gap |

Q 32 The Fermi level of N type of semiconductor at 0 K is

- | | |
|--|---|
| a In between bottom of conduction band
at bottom of donor energy levels | b In between the bottom of the donor
energy levels and the middle of the
band gap |
| c At the center of the band gap | d Near acceptor levels |

Q 33 The Fermi level of P type of semiconductor at 0 K is

- | | |
|---|--|
| a At the center of the band gap | b In between the top of the acceptor
levels and the top of the valance band |
| c In between the bottom of the donor
energy levels and the middle of the band
gap | d Near donor levels |

Q 34 The Fermi level of N type of semiconductor at 300 K is

- | | |
|---|--|
| a At the center of the band gap | b In between the top of the acceptor
levels and middle of the band gap |
| c In between the bottom of the donor
energy levels and the middle of the band
gap | d In between the top of the acceptor
levels and the top of the valance band |

Q 35 The Fermi level of P type of semiconductor at 300 K is

- a At the center of the band gap
- b In between the top of the acceptor levels and middle of the band gap
- c In between the bottom of the donor energy levels and the middle of the band gap
- d In between the top of the acceptor levels and the top of the valance band gap

Q 36 In N type of semiconductor, when the concentration of the pentavalent dopant increases, the Fermi level

- a Enters in the valance band
- b Enters in the conduction band
- c Shifts towards the middle of the band gap
- d Shifts away from the middle of the band gap

Q 37 In P type of semiconductor, when the concentration of the trivalent dopant increases, the Fermi level

- a Enters in the valance band
- b Enters in the conduction band
- c Shifts towards the middle of the band gap
- d Shifts away from the middle of the band gap

Q38 Solar Photovoltaic cell converts light in to

- a Heat
- b Electrical current
- c Electrical voltage
- d Electrical power

Q 39 The effective mass of an electron is given by

a $m^* = \frac{\hbar^2}{\frac{d^2E}{dk^2}}$	b $m^* = \frac{\hbar^2}{\frac{d^2E}{dx^2}}$
c $m^* = \frac{\hbar^2}{\frac{d^2E}{dy^2}}$	d $m^* = \frac{\hbar^2}{\frac{d^2E}{dt^2}}$

Q40 In an unbiased PN junction diode

- a The Fermi levels of N side and P side are at the same level
- b The Fermi levels of the N side are near the conduction band and the Fermi levels of P side are near the valance band
- c The Fermi levels of the P side are near the conduction band and Fermi levels of N side are near the valance band
- d The Fermi levels of the N side are in the conduction band the Fermi levels of the P side are in the valance band

REFERENCE BOOKS

1. Fundamentals of Physics, Halliday, Resnick, Wiley
2. Concepts of Modern Physics, Arthur Beiser, tata McGraw Hill
3. A textbook of Engineering Physics, Arthur Beiser
4. Solid State Physics, N. W. Ashcroft and N. D. Mermin, (CBS Publishing Asia Ltd.)
5. Introduction to Solid State Physics, Charles Kittel, (John Wiley and Sons.)
6. Introductory Solid State Physics, H. P. Myers, (Viva Books Pvt. Ltd.)
7. Solid State Physics, H. Ibach and H. Luth, (springer-Verlag).
8. Fundamentals of Solid State Physics, J. R. Christman, (John Wiley and Sons.)
9. Solid State Physics, A. J. Dekkar, (Prentice Hall). 7. Physics of Semiconductor Devices, S. M. Sze (John Wiley and Sons.)
10. Physics of Semiconductor Devices – S.M. Sze
11. Physics Solid State Devices – Streetman B.B.
12. Semiconductor Physics – Smith
13. Fundamentals of Semiconductor Devices – J. Lindmayer and C.Y. Wrigley
14. Physics of Semiconductor Devices – Michael shur
15. Introduction to Semiconductor devices – K.J.M. Rao

WORLD WIDE WEB

1. www.electronics-tutorials.ws
2. www.hyperphysics.com

CHAPTER 10

Superconductivity



The photograph on the left shows a superconducting specimen floating over a magnet. This is due to Meissner effect, which is invariably exhibited by all superconductors. The photograph on the right shows one of the fastest railway trains in the world, named maglev. It's speed is 501 km/h. Maglev means magnetically levitated train; and it is based on Meissner effect. What is Meissner effect? How it is exhibited by superconductors? What is superconductivity?

The answer to these questions are in this chapter

Index

10.1 INTRODUCTION

The resistance of some materials falls to zero below critical temperature

10.2 SUPERCONDUCTING MATERIALS: A HISTORICAL PERSPECTIVE

Metals, alloys and ceramics; from $T_c = 0.015\text{ K}$ to 125 K

10.3 PROPERTIES OF SUPERCONDUCTORS

Why superconductivity is considered to be a mysterious phenomenon

10.4 MEISSNER EFFECT

Superconductors are perfectly diamagnetic

10.5 EFFECT OF MAGNETIC FIELD ON SUPERCONDUCTIVITY

Superconductivity is destroyed above critical magnetic field

10.6 TYPE I AND TYPE II SUPERCONDUCTORS

The classification is based on critical magnetic field

10.7 HIGH TEMPERATURE SUPERCONDUCTORS:

Their critical temperature can be easily achieved by using liquid nitrogen

10.8 BCS THEORY

Ordered flow of Cooper pairs

10.9 APPLICATIONS OF SUPERCONDUCTIVITY

From ultrathin SQUIDs to giant MAGLEVs

10.10 JOSEPHSON EFFECT

Tunneling of Cooper pairs through a thin insulating barrier



Heike Kamerlingh Onnes (1853-1926): Heike Kamerlingh Onnes obtained his doctorate degree in 1879. He was appointed as a professor in Physics in 1882 in Leiden University, Netherlands. He worked under Kirchoff. In Leyden university, he established his own laboratory, where he studied physics of materials at low temperatures. The smallest temperature that he achieved in his lab was 0.9 °K. At those times his lab was recognized as ‘coldest spot’ on the earth! The other investigations which took place in Onne’s lab were related to thermodynamics, radioactive law , observations on optical, magnetic and electrical phenomena (fluorescence, phosphorescence etc.), magnetic rotation of the polarization plane, absorption spectra of crystals in the magnetic field; also the Hall effect, dielectric constants, and especially the resistance of metals at low temperatures. A momentous discovery (1911) was that of the *superconductivity* in pure metals such as mercury, tin and lead at very low temperatures. Many foreign Physicists came to Leyden to work in Onne’s laboratory for shorter or longer periods. At the early age of 30, Onnes was appointed as a member of Royal Academy of Sciences. Throughout his career, Onnes won several awards and honors, one of which was Nobel prize, which was awarded to him in 1913 for his investigations on the properties of matter at low temperatures.

10.1 INTRODUCTION

The resistance of some materials falls to zero below critical temperature

Heike Kamerlingh Onnes, a Dutch Physicist was the first to liquefy helium at the temperature of 4.2 K in 1908. He was interested in studying the Physics of materials at low temperatures. In 1911, he found that electrical resistance of mercury abruptly fell to zero when it was cooled below 4.2 K. (Fig 10.1). He called this phenomenon as superconductivity. Soon, it was established that the resistance of some metals, alloys and ceramics falls to zero when they are cooled below critical temperature. Superconductivity, which is yet a low temperature phenomenon, appears to be quite novel, fascinating and miraculous due to number of reasons. One is that, the elements such as gold, silver and copper which are good conductors of electricity at room temperature, do not exhibit superconductivity at all. However, some ceramics, which are insulators at room temperature, exhibit superconductivity below critical temperature. Further, though superconductors have zero resistance, they are not just ‘ideal’ conductors, in the sense that, in addition to their zero electrical resistance, they also exhibit perfect diamagnetism. Superconductivity remains is one of the most widely researched phenomena in the world and

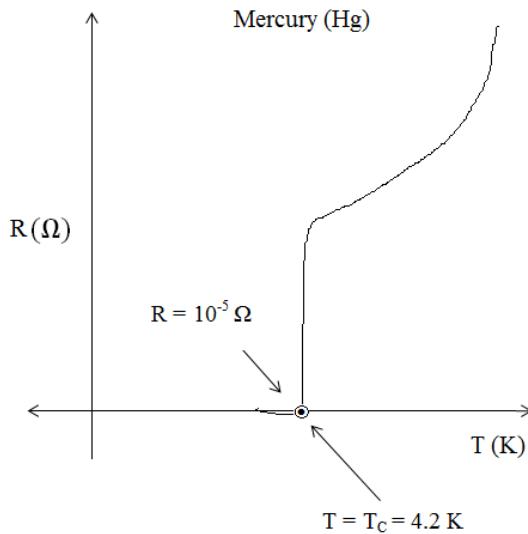


Figure 10.1 Superconductivity in Mercury (Hg).

(For chemically pure and defect free materials, the normal to superconducting state is sharp, while for materials having defects and impurities the transition is broad up to $1/10^{\text{th}}$ of a degree)

despite that, the room temperature superconductor is yet to be discovered! However, the low temperature superconductors have acquired their own place in technology. Maglevs, SQUIDS, superconducting high power electromagnets, solenoids, transformers, motors, generators etc are a few examples. Superconductivity has a potential of many other applications in the future, one of which can be superconducting power transmission with zero I^2R losses. This chapter is aimed at the discussion of the theory and applications of superconductors.

Fig (10.1 and 10.2) show the difference between the superconductors and the normal conductors. The resistance of conductors is due to disorders, thermal vibrations and electron scattering due to impurities and defects. It can be noted from the Fig 10.2 that the resistance of normal conductors decreases with decrease in the temperature.

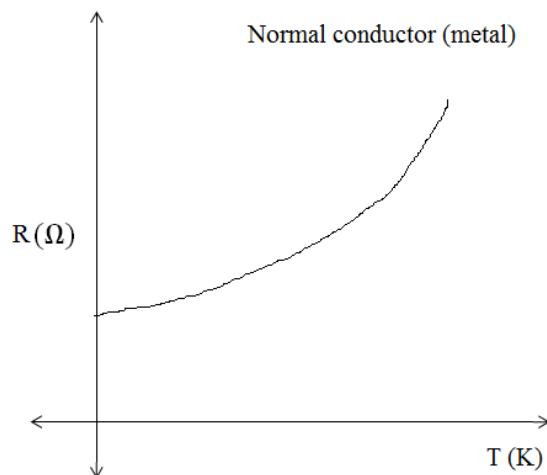


Figure (10.2): How resistance of normal conductors varies with temperature

However the metals with impurities and defects may have some resistance even at absolute zero. In contrast, the resistance of superconductors falls to absolute zero at non-zero temperatures. The transition from normal state to superconducting state is sharp only for pure metals

10.2 SUPERCONDUCTING MATERIALS: A HISTORICAL PERSPECTIVE

Metals, alloys and ceramics; from $T_c = 0.015\text{ K}$ to 125 K

In 1911, the first material in which superconductivity was discovered by Heike Kamerlingh Onnes was mercury. It's critical temperature was recorded to be 4.15 K. At those times, liquid helium was available only in Onnes' lab and thus, after 1911, in Onnes' lab itself, the superconductivity was discovered in three materials, namely thallium ($T_c = 2.4\text{ K}$), indium ($T_c = 3.4\text{ K}$), tin ($T_c = 3.7\text{ K}$), lead ($T_c = 7.2\text{ K}$). Later on in 1925, Walther Meissner discovered superconductivity in niobium ($T_c = 9.2\text{ K}$). At those times $T_c = 9.2\text{ K}$ was the highest critical temperature achieved. Niobium still remains the an elemental superconductor with highest critical temperature. Till 1940 it was realized that metallic compounds could have still higher critical temperatures. By 1940 it was found that NbC has a critical temperature of 10.1 K. In 1954 it was Nb₃Sn with critical temperature of 23.2 K. In 1973 it was found that Nb₃Ge has a critical temperature of 23.2 K. This achievement remained at the top till 1986. In 1986 Muller and Bendroz (Nobel laureates) demonstrated superconductivity in copper oxides. It was found that lanthanum barium copper oxide has a critical temperature of 35 K. Later on it was shown that yttrium barium copper oxide has a critical temperature of 91 K. Still later bismuth and thallium oxides were found with T_c values up to 125 K. Such superconductors are called as high temperature superconductors.

Superconductivity is exhibited by many metals. There are around 26 elements which exhibit superconductivity at lower temperatures. Thousands of alloys are known to exhibit the superconductivity. It is found that the superconductors, which have zero resistance below critical temperature, are the poor conductors of electricity at room temperatures, while the elements such

Sr. No.	Material	Critical temperature with T_c in K
1	Copper, silver, gold	Non-superconducting
2	Rhodium	240×10^{-6}
3	Aluminum	1.1
4	Tin	3.72
5	Mercury	4.15
6	Lead	7.2
7	Niobium	9.3
8	Niobium titanium alloys	9-11
9	Lead-Molybdenum sulfide	14
10	Niobium-tin	18.3
11	Vanadium gallium	15.4
12	Niobium germanium	23.3

Table (10.1) Examples of some superconductors

as silver, gold and copper, which are the good conductors of electricity at room temperatures, do not exhibit superconductivity at all. Further, iron, cobalt and nickel which are known to be ferromagnetic, do not exhibit superconductivity. The elements such as silicon, germanium, which are semiconductors at room temperatures, show superconductivity at low temperatures when they are kept under pressure. Amongst the elements, niobium has highest critical temperature (9.46 K) and tungsten has lowest critical temperature (0.015 K). It is interesting to note that CuS is a superconductor, but Cu and S are not. Also, Au₃Bi exhibit superconductivity but Au and Bi do not. The table 10.1 gives a tentative list of superconductors with their critical temperatures.

Other factors which affect superconductivity:

- i. **Stress:** When stress is applied, the dimensions of the material change. Hence critical temperature of the superconductor increases when stress is applied
- ii. **Impurities:** Almost all the properties, especially the magnetic properties are altered due to impurities
- iii. **Size:** The critical magnetic field, magnetic permeability and other properties change, particularly when the size is reduced below 100 μm
- iv. **Frequency:** For very large frequencies ($>$ GHz), the superconductors offer resistance even if $T < T_c$

10.3 CHARACTERISTICS OF SUPERCONDUCTORS:

Why superconductivity is considered to be a mysterious phenomenon

	Walther Meissner (1882–1974): He was a German Physicist who received his education in Technical University, Berlin, where his Ph. D. guide was Max Planck, a Nobel laureate. In 1922-25 Meissner designed one of the world's largest helium liquefier, and thereby in 1933, he along with Robert Ochsenfeld, discovered an effect now known by his name. Meissner effect relates to damping of the magnetic field in superconductors. Thereafter he worked in Technical University, Munich. He lived up to an age 92, but for last several years, he lived alone till his death in 1974
---	--

Meissner effect:

Superconductors are perfectly diamagnetic

Consider a superconducting coil having no current passing through it. If such coil is brought near the magnetic field, then there will be $\frac{d\phi}{dt}$, which will generate induced emf. Due to induced emf there will be induced current inside the coil which will produce opposite magnetic field. Now if the motion of the coil is stopped then induced emf will become zero, however the current and hence the opposite magnetic field will still persist due to zero resistance. Due to this opposite magnetic field superconductors exhibit perfect diamagnetism. According to Meissner effect,

superconductors expel the magnetic flux. Magnetic flux is not allowed to pass through the superconductor. The magnetic field inside a superconductor is always zero.

Now, consider a superconductor outside magnetic field (H). If such superconductor is brought inside the magnetic field then the magnetic field is expelled from the superconductor. This is because, in presence of external magnetic field H , the superconductor generates its own surface currents, so that an opposite magnetic field ($-M$) is generated. The net magnetic field through the superconductor is zero. If the external magnetic field is switched off, the internal reverse magnetic field is also reduced to zero.

Meissner effect, Path I

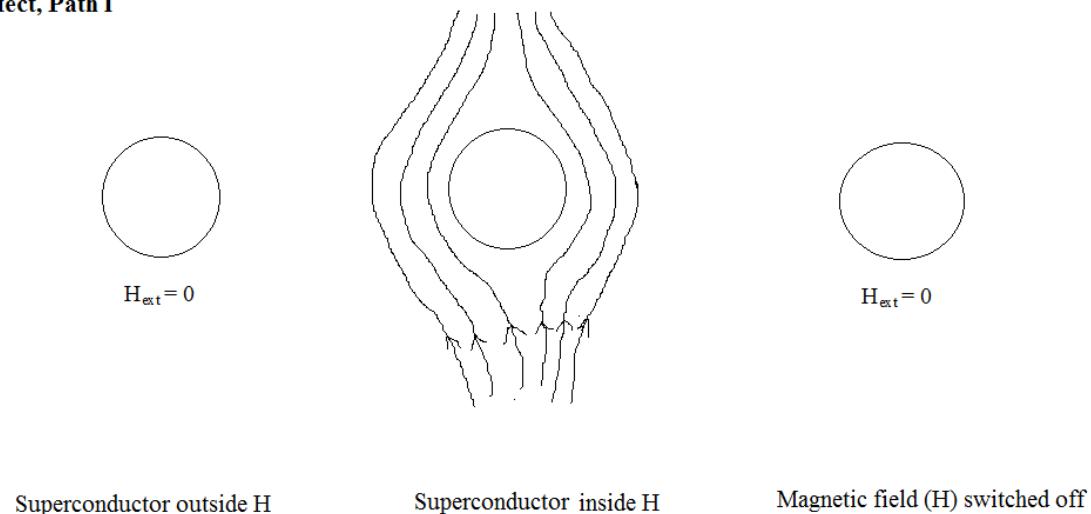


Figure 10.3 Meissner effect: Path I

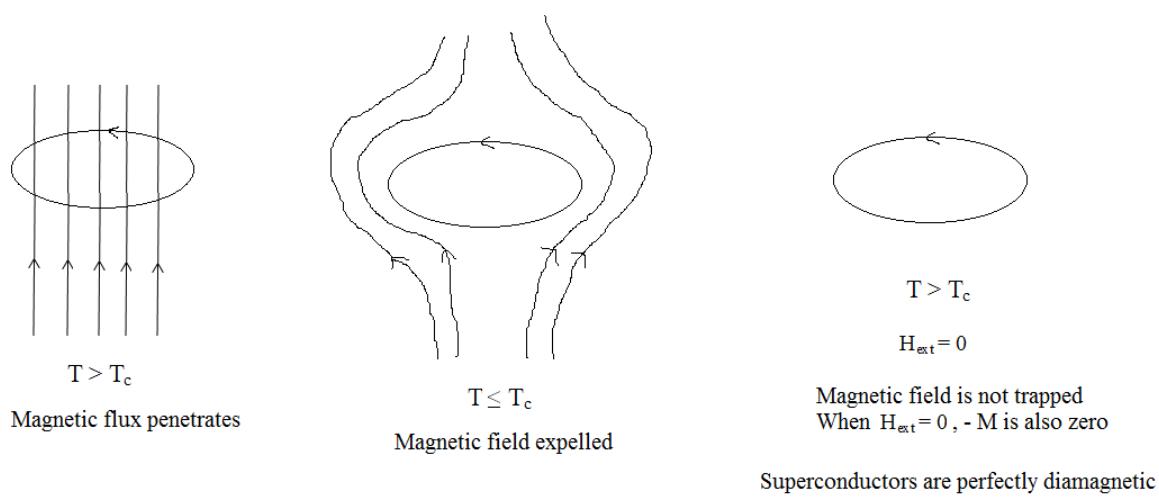


Figure (10.4) Meissner effect: Path II

Now, consider Fig. (10.4). When the temperature of the superconductor is above its critical value, the magnetic field lines pass through the superconductor. If the temperature of the

superconductor is decreased below critical value, the specimen acquires a superconducting state, and thus the magnetic field is expelled. The material in the superconducting state generates its own surface currents, due to which an opposite internal magnetic field ($-M$) is generated. The net magnetic flux through any material in it's superconducting state is always zero. If the external magnetic field (when the material is in superconducting state) is switched off, or if the temperature exceeds critical temperature, specimen loses superconducting properties. Then internal reverse magnetic field of the superconductor also reduces to zero.

Consider a superconductor being held in magnetic field, H . As superconductors expel all the magnetic flux, the net magnetic field, B inside the superconductor is zero. We have

$$B = \mu_0(M + H)$$

$$0 = \mu_0(M + H)$$

Thus

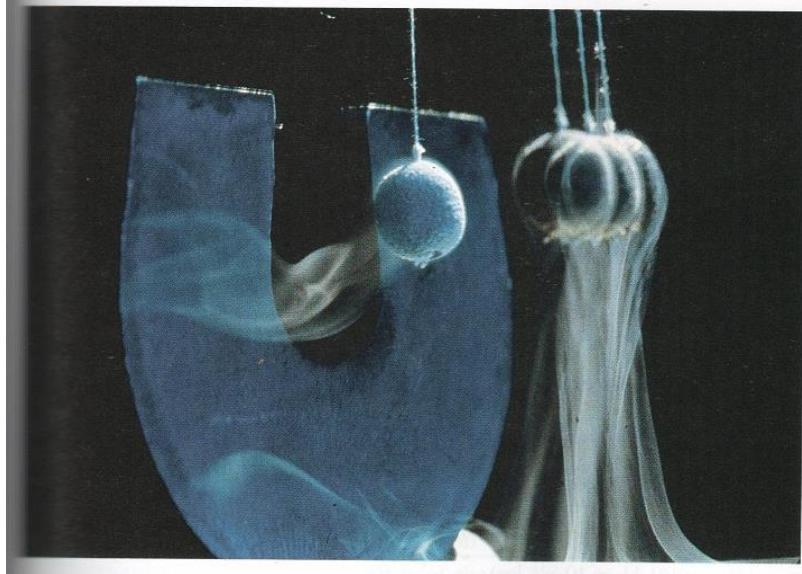
$$-M = H \quad \dots(10.1)$$

This indicates that when placed in external magnetic field, H , the superconducting specimen generates it's own internal reverse magnetic field ($-M$), such that the net magnetic induction B inside the specimen is zero

Magnetic susceptibility is given by

$$\chi = \frac{-M}{H}$$

$$\Rightarrow \chi = -1 \quad \dots(10.2)$$



A superconductor is a perfectly diamagnetic. Here, the superconducting pendulum bob is repelled by the permanent magnet

This indicates that superconductors are perfectly diamagnetic. Thus in superconductor , we don't have ($R = 0$), but $R = 0$ and $\chi = -1$.

As we shall discuss later in details

1. All superconductors are not perfectly diamagnetic. For type II superconductors, the magnetic flux can penetrate through the specimen.
2. To some extent, the magnetic flux permeates through all the superconductors. The extent to which the magnetic flux permeates the superconducting specimen is called as London penetration depth.

One of the most promising application of this effect, is a MAGLEV (magnetically levitated train)

Difference between an ideal conductor and superconductor:

For an ideal conductor, we have $R = 0$ and hence $\rho = 0$

We also have $E = J\rho$. As $\rho = 0$, $E = 0$, thus the electric field inside an ideal conductor is zero

According to Maxwell's equation, we have

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

As the electric field (E) inside the ideal conductor is zero,

$$-\frac{\partial B}{\partial t} = 0$$

$$\Rightarrow B = \text{constant}$$

Thus the magnetic field inside the ideal conductor is constant (not zero). This indicates that in ideal conductor, the magnetic field is frozen. Thus if an ideal conductor is brought in an external

Superconductor	Ideal conductor
$R=0$	$R=0$
$\chi = -1$	$\chi \neq -1$
$\frac{\partial B}{\partial t} = 0$	$\frac{\partial B}{\partial t} = 0$
$B = 0$	$B = \text{Constant (Freezed)}$

Table (10.2) Difference between a superconductor and an ideal conductor

magnetic field, then irrespective of the value of external magnetic field (zero, or finite), B inside the ideal conductor is always constant. However, in superconductors, for a given value of external magnetic field, H , there exists a reverse magnetic field $-M$, such that, $-M = H$. Thus the net magnetic field, B inside the superconductor is always zero. If an ideal conductor is kept in external magnetic field, H , then it will contain a magnetic flux, B ($= \text{constant}$), depending upon its permeability. When external magnetic field is switched off, the magnetic induction inside the ideal conductor does not become zero, but remains constant. The frozen/trapped magnetic field inside an ideal conductor does not increase or decrease on changing the external magnetic field. An ideal conductor traps the magnetic field, while a superconductor expels the

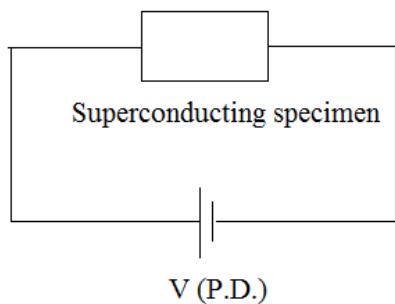
magnetic field completely. Table (10.2) gives a comparison between an ideal conductor and a superconductor.

Persistent currents: Consider a superconducting coil brought near a magnet. As the coil approaches the magnet, there is $\frac{d\phi}{dt}$ and hence an induced emf (ε). Due to induced emf, there is induced current. Now if the motion of the superconductor is stopped, the $\frac{d\phi}{dt}$ and hence an induced emf, ε , reduces to zero. However, as the resistance of superconductor is zero, the induced current still persists for a longer period without any reduction. Such currents are called as persistent currents. It has been shown that, in superconductors, the current persists without any reduction for several years. Persistent current can also be obtained by decreasing the temperature below its critical value, while current is flowing through the superconductor. Another way of obtaining persistent current is to keep the superconducting coil ($T < T_c$) in magnetic field and then switching off the magnetic field. The sudden decrease in the magnetic field gives the required emf. Persistent current is an important feature of all superconductors. The superconducting coils carrying persistent currents can be used as steady and un-diminishing source of magnetic field.

How to test superconductivity. The superconducting state of a material can be verified by two methods

Method I: Consider a superconducting specimen (Fig. 10.5) applied with a battery having P.D. of V . If the temperature of the specimen is above the critical temperature, then the specimen has finite nonzero resistance, R . In such cases the P.D. measured across the specimen is finite (V). When the temperature of the specimen is reduced to less than or equal to critical temperature, the resistance of the specimen reduces to zero. According to Ohm's law ($V = IR$), the P.D. falls to zero. Thus if the P.D. across the specimen is zero, in spite of application of battery, then the specimen can be considered to be in superconducting state

$$T > T_c, R \neq 0, V (\text{P.D.}) \neq 0$$



$$T \leq T_c, R = 0, V (\text{P.D.}) = 0$$

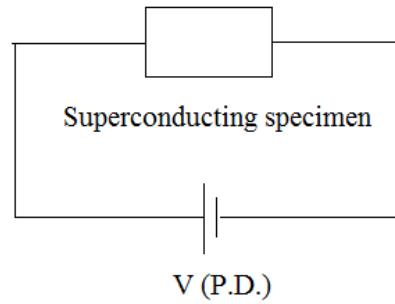
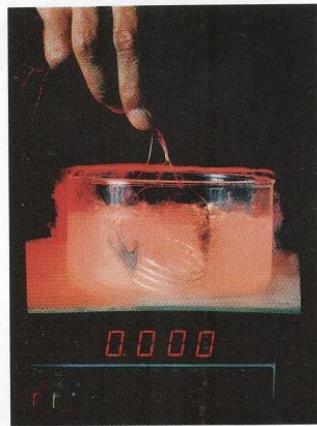


Figure (10.5) Testing superconductivity: Method I



An Ohmmeter measures zero resistance across this superconducting coil, which is made of a barium-yttrium-copper oxide compound and cooled in liquid nitrogen

Method II: Consider a superconducting coil. If a magnet is brought close to the coil and then taken away, then according to Lenz's law, a current is induced in the superconducting coil. The superconductor has zero resistance, therefore the induced current does not decrease. This current is a persistent current. As the current does not decrease, there is no $\frac{d\phi}{dt}$ around the superconductor. Thus there is no emf induced in a test coil kept near the superconductor. Thus there is no current across the test coil. The absence of induced current in test coil over a considerable period indicates that the coil is in superconducting state.

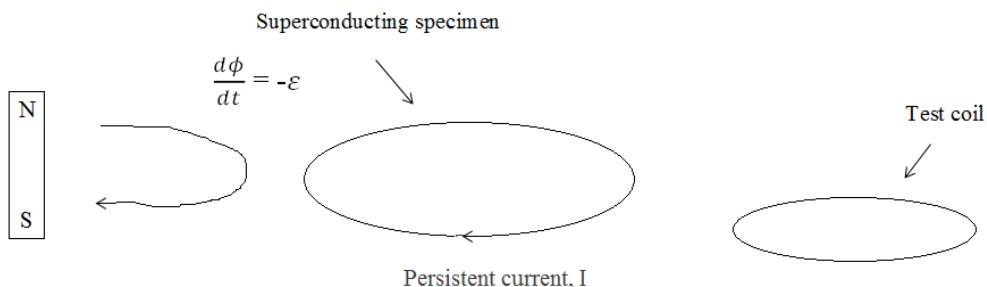


Figure 10.6 Testing superconductivity: Method II

Three natural enemies of superconductivity: Superconductivity has three natural enemies; one is temperature, second is magnetic field and third is the current density. The material is said to be in superconducting state when $T < T_c$, $H < H_c$ and $J < J_c$. As we shall see further, superconductivity is based on formation of Cooper pairs of electrons. The cooper pairs are broken when $T > T_c$. The Cooper pair contains electrons with opposite spins. When $H > H_c$, the spins of the electrons are aligned and thus Cooper pairs are broken and superconductivity is destroyed. Superconductor generates its own magnetic field due to passage of a current through

it. When $J > J_c$ the superconductor's own magnetic field exceeds H_c and thus the superconductivity is destroyed.

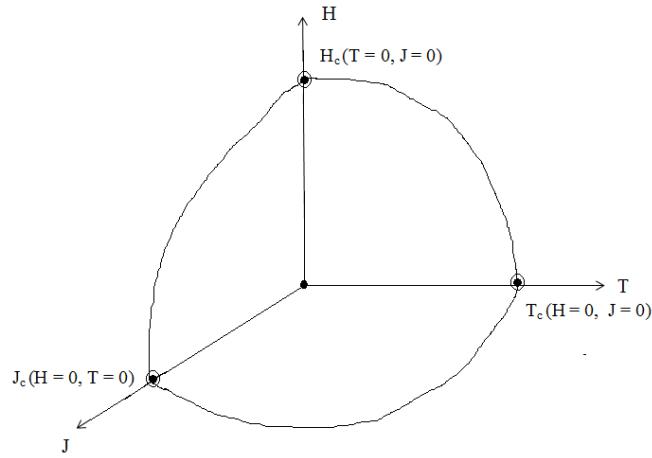


Figure 10.6: Three enemies of superconductivity

10.4 EFFECT OF MAGNETIC FIELD ON SUPERCONDUCTIVITY

Superconductivity is destroyed above critical magnetic field

As we shall see later, superconducting state is due to bound electron pairs called Cooper pairs. The electrons in Cooper pair have opposite (anti-parallel) spins. If superconductor is kept in magnetic field and if magnetic field is increased, then the magnetic field, at it's critical value (H_c) aligns the spins of all the electrons in cooper pairs in the same direction. Thus cooper pairs are broken and superconducting state is lost. The magnetic field at which the superconductivity is lost is called as critical magnetic field (H_c). Fig. (10.7) and Eqn. (10.3) shows the dependence of the critical magnetic field on the temperature.

$$H_c(T) = H_c(0) \left[1 - \left(\frac{T}{T_c} \right)^2 \right] \quad \dots(10.3)$$

Fig. (10.7) and Eqn (10.3) indicate that at $T = 0$, the critical magnetic field is maximum and is given by $H_c(0)$. As T increases, the ratio T/T_c increases due to which $\left[1 - \frac{T}{T_c} \right]$ decreases and thus $H_c(T)$ also decreases, when $T = T_c$, H_c reduces to zero. This indicates that when temperature increases, the critical magnetic field required to destroy superconductivity also decreases.

In Fig (10.7), consider point P. It is within the curve. The specimen is then in superconducting state ($R = 0$) and $\chi = -1$. However, if we proceed away along any direction (T or H_c), and if we reach beyond the curve, then the specimen is transformed to an ordinary (normal state).

The critical magnetic field depends upon the material. Table (10.3) shows the critical magnetic fields of some Type I superconductors. It can be noted that the critical magnetic fields of all the type I superconductors are less than 0.2 T. Therefore, such superconductors are called as soft superconductors. They cannot be used in the situations where magnetic fields are high. As

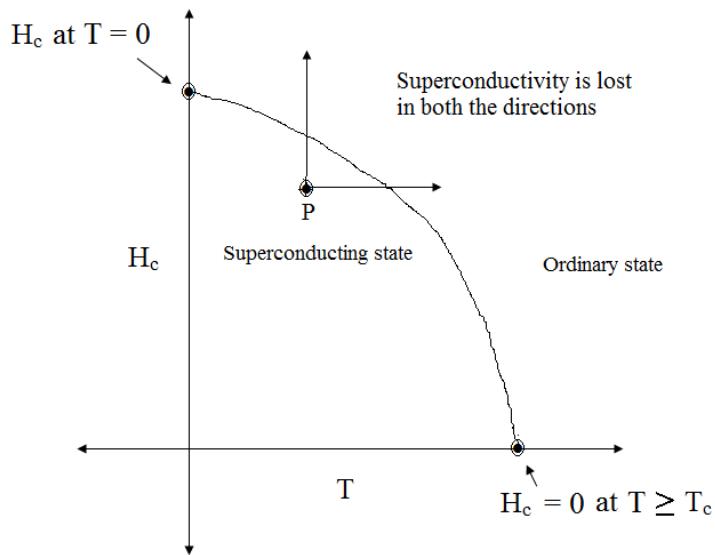


Figure (10.7): Effect of magnetic field on superconductivity

we shall see later, the critical magnetic fields of Type II superconductors (Hard superconductors) are considerably high.

Superconductor	Critical temperature (T_c), K	Critical magnetic field $H_c(0)$, T
Tungsten (W)	0.015	0.0001
Titanium (Ti)	0.390	0.01
Cadmium (Cd)	0.560	0.003
Zinc (Zn)	0.85	0.0054
Molybdenum (Mo)	0.920	0.0095
(Aluminum) Al	1.18	0.0105
Indium (In)	3.408	0.0281
Tin (Sn)	3.722	0.0305
Mercury (Hg)	4.153	0.0411
Vanadium (V)	5.380	0.1420
Lead (Pb)	7.193	0.0803
Niobium (Nb)	9.460	0.1980

Table (10.3) Critical magnetic fields of some Type I superconductors

Example (10.1)

The critical temperature of mercury is 4.153 K. The maximum critical magnetic field at T = 0 K is 0.0411 T. Calculate the critical magnetic field at the temperature 3 K.

Solution:

We have

$$H_c(T) = H_c(0) \left[1 - \left(\frac{T}{T_c} \right)^2 \right]$$

T_c = 4.153 K,

H_c(0) = 0.0411 Tesla

T = 3 K

Substituting

$$H_c(T) = 0.0411 \times \left[1 - \left(\frac{3}{4.153} \right)^2 \right]$$

$$H_c(3 K) = 0.01965 K$$

Example (10.2):

The transition temperature for Pb is 7.2 K. However at 5 K, it loses the superconducting property, if subjected to a magnetic field of 3.3×10^4 A/m. Find the maximum value of H which will allow the metal to retain its superconductivity at 0 K.

Solution:

We have

$$H_c(T) = H_c(0) \left[1 - \left(\frac{T}{T_c} \right)^2 \right]$$

Given

T_c = 7.2 K, T = 5 K, H_c(T) = H_c(5 K) = 3.3×10^4 A/m, H_c(0) = ?

Substituting

$$3.3 \times 10^4 = H_c(0) \times \left[1 - \left(\frac{5}{7.2} \right)^2 \right]$$

$$H_c(0) = 6.37 \times 10^3 \text{ A/m}$$

Example (10.3):

The transition temperature for lead is 7.26 K. The maximum critical field for the material is $8 \times 10^5 \text{ A/m}$. Lead has to be used as a superconductor subjected to a magnetic field of $4 \times 10^4 \text{ A/m}$. Calculate the temperature below which the material has to be kept.

Solution:

We have

$$H_c(T) = H_c(0) \left[1 - \left(\frac{T}{T_c} \right)^2 \right]$$

Thus

$$T = T_c \times \sqrt{1 - \frac{H_c(T)}{H_c(0)}}$$

Substituting

$$T = 7.26 \times \sqrt{1 - \frac{4 \times 10^4}{8 \times 10^5}}$$

$$T = 7.08 \text{ K}$$

Critical current density (J_c) . The current passing through the superconductor generates its own magnetic field. When the current density reaches its critical value, the corresponding magnetic field destroys superconductivity. This effect is called as Silsbee effect, which was discovered in 1916

$$I_c = 2\pi R H_c$$

10.5 TYPE I AND TYPE II SUPERCONDUCTORS

The classification is based on critical magnetic fields

Till now we have seen that the magnetism and superconductivity are natural enemies. We have also seen that superconductors are perfectly diamagnetic, they do not allow the magnetic field to

permeate through them. The net magnetic field inside the superconductors is always zero. This is because, when placed in external magnetic field, the superconductor generates its own reverse magnetic field ($-M$) which cancels the external magnetic field (H). Consider Fig (10.8). It is observed that as H increases, $-M$ also increases linearly. When H reaches its critical value, $-M$ suddenly drops to zero, the superconductivity is destroyed and the magnetic flux permeates through the material. The transition from superconducting to normal state is sudden. It can also be seen that the graph of $-M$ Vs H is linear and inclined at 45° . It can also be observed that at $H = H_c$, the graph suddenly drops to zero. As we have seen, this is Meissner effect, and such superconductors are called as Type I superconductors. As it can be noted from Table (10.3), the critical magnetic field of Type I superconductors is considerably low. Therefore these are also called as soft superconductors. Some typical examples of type I superconductors are aluminum, lead, indium etc.. The critical magnetic fields of type I superconductors are in the range of 0.01 T to 0.2 T. The London penetration depth (to be discussed) of type I superconductor is small. The currents that tend to cancel the external magnetic field are along surface only.

In 1935, Rjabinin and Lev Shubnikov discovered Type-II superconductors experimentally. Though the thermal behavior of Type I and Type II superconductors is same, the magnetic behavior (especially Meissner effect) of type II superconductors is completely different when compared to Type I superconductors.

Consider Fig (10.9). We observe that, as the external magnetic field (H) increases, the reverse magnetic field ($-M$) inside the superconductor also increases, thus the net magnetic field inside the superconductor remains zero (Meissner effect). However, when H exceeds H_{c1} , the magnetic behavior of Type II superconductor becomes different than type I. The magnetic field starts penetrating through the specimen. As H increases, increasingly greater magnetic field penetrates through the specimen. However, in this region, the electrical properties of type II

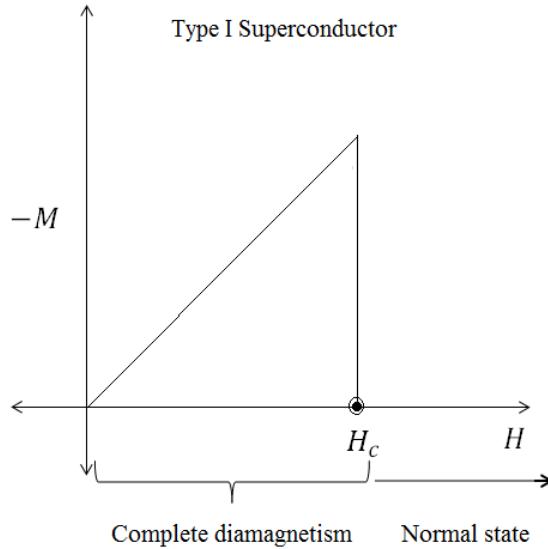


Figure (10.8) Magnetic behavior of Type I Superconductors

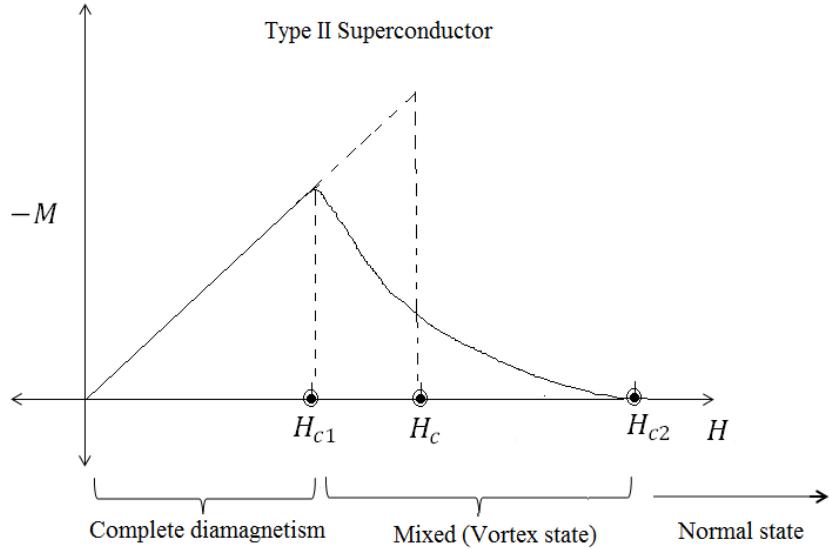


Figure (10.9): The magnetic behavior of type II superconductors

superconductors are still similar to type I; and thus the resistance of type II remains zero. As the H exceeds H_{c2} , the type II superconductor behaves as an ordinary conductor. Thus all the magnetic flux penetrates through the superconductor and the resistance of the superconductor becomes nonzero. Thus type II superconductors have two critical magnetic fields, the H_{c1} and H_{c2} . The transition of properties of type II superconductors from H_{c1} to H_{c2} is gradual and smooth. The typical value of H_{c2} for type II superconductors is 20 T.

Thus it can be seen that the behavior of the type I and type II superconductors below $H = H_{c1}$ and above $H = H_{c2}$ is same. Below H_{c1} , both the superconductors have zero resistance and are perfectly diamagnetic; no magnetic field passes through them. The resistance of both the superconductors is zero when $H \leq H_{c1}$. When $H \geq H_{c2}$, the type II superconductor behaves just like an ordinary conductor. In the region $H_{c1} < H < H_{c2}$, type II superconductor is in magnetically mixed state, however, the resistance is zero. (refer Fig 10.10 and 10.11). This state is also called as Vortex state. The magnetic flux in this region is quantized. It increases with H , in the steps of $\frac{h}{2e}$. A quantum of magnetic flux passing through the superconductor in the mixed state is called as fluxon or fluxoid. As H increases in the region in between H_{c1} and H_{c2} the number of flux quanta (fluxons) also increases. When H exceeds H_{c2} , the entire specimen is filled with fluxons. We have noted that in the mixed state, the resistance of the superconductor is zero. Thus in the mixed state, the superconductor contains filaments surrounded by vortices which allow the current to pass through them. The magnetic flux passes through the filaments, while the current passes through surrounding vortices. Thus in the region H_{c1} to H_{c2} the Meissner effect is only partially obeyed. In this region, the type II superconductors are not perfectly diamagnetic.

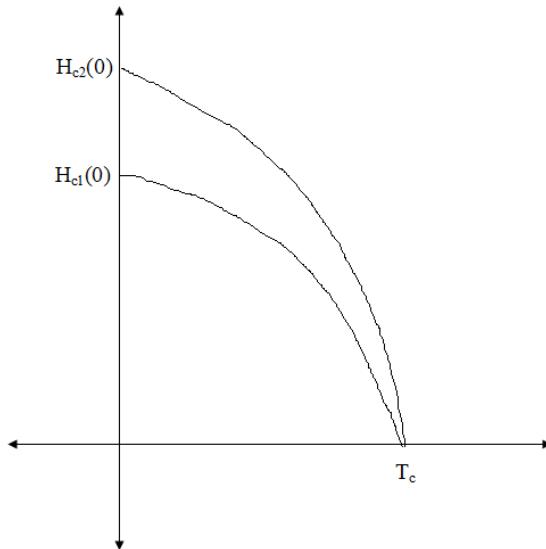


Figure (10.10): Dependence of H_{c1} and H_{c2} on temperature

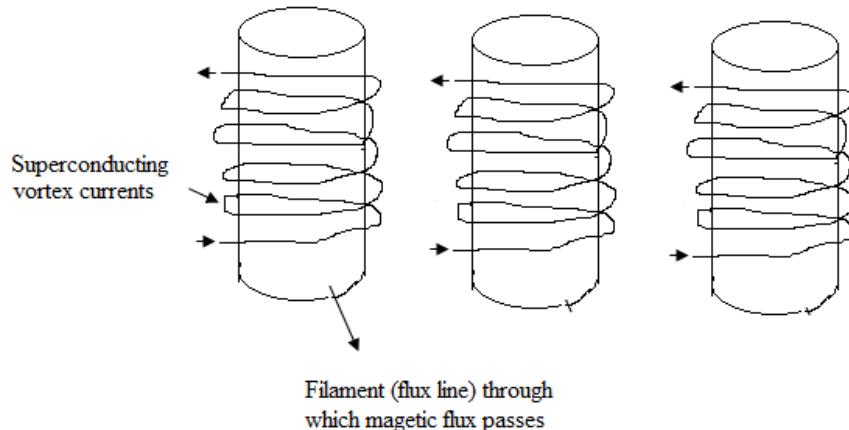


Figure (10.11): Filaments and vortices in type II superconductors

The values of H_{c2} for type II superconductors are high. Thus these superconductors are used to generate heavy magnetic fields by using electromagnets and solenoids. The heavy magnetic fields generated using type II superconductors are used in MRI, nuclear fusion, accelerators and Maglevs. Due to greater critical magnetic fields (H_{c2}), type II superconductors are also called as Hard superconductors. Table (10.3) and (10.4) gives the examples of type I and type II superconductors. Two typical examples of type II superconductors are Nb-Ti ($H_{c2} = 13$ T) and Nb₃Sn ($H_{c2} = 20$ T). It can be noted that type I superconductors are elements, while type II superconductors are alloys and ceramics. As we shall see later, all high temperature superconductors are type II superconductors. The London penetration depth of type II superconductor is large as compared to type I. Unlike type I, the currents in the superconductors are along surface as well as in volume also.

Superconductor	Critical temperature, (T_c)K	Critical magnetic field, $H_c(0)$, T
Nb ₃ Ti	10.0	15.7
Nb ₃ Sn <i>(Most important commercial alloy)</i>	18	24.5
B ₃ Al	18.7	32.4
Nb ₃ Ge	28.2	38.0
Nb ₃ Al	18.7	32.4
Nb ₃ (AlGe)	20.7	44
V ₃ Ge	14.8	2.08
V ₃ Si	16.9	2.35
NbTi	10	13
PbMoS	14.4	6.0
High temperature (Type II) superconductors <i>(all high temperature superconductors are type II superconductors)</i>		
La ₁ Ba ₂ Cu ₃ O ₇	30	-
LaBaCuO	40	-
Bi ₂ Sr ₂ CaCu ₂ O ₈	85	-
YBa ₂ Cu ₃ O ₇	92	-
HgBa ₂ CuO ₄	94	-
YBa ₂ CuO ₇	95	-
Bi ₂ Sr ₂ Ca ₂ Cu _{10+δ}	105	-
Bi ₂ Sr ₂ Ca ₂ Cu ₃ O ₆	110	-
Th ₂ CaBa ₂ Cu ₂ O _{10+δ}	125	-
HgBa ₂ CaCu ₂ O ₆	128	-
HgBa ₂ Ca ₂ Cu ₃ O ₈	134	-

Table (10.4) Critical magnetic fields of some Type II superconductors

10.6 HIGH TEMPERATURE SUPERCONDUCTORS:

Their critical temperature can be easily achieved by using liquid nitrogen

Till 1986, the highest critical temperature known was 27 K. The superconductors available in those times were metals and alloys. Physicists had almost concluded that it was impossible to find a high temperature superconductor. However, in 1986 , two Nobel laureates Alex Muller, George Bendroz (Switzerland), crossed this limit by discovering ceramic superconductors. Generally ceramics are insulators at room temperature; however it is shown that, below critical temperature, they show superconductivity (zero resistance and Meissner effect). The first ceramic in which superconductivity was discovered was lanthanum barium and copper oxide having critical temperature, (T_c) as 35 K. After this discovery, there was a worldwide research in this field due to which number of other ceramic superconductors were discovered. One of them was developed by Wu and Chu and their collaborators. It was yttrium barium copper oxide, which had a critical temperature of 91 K. These were oxides of mercury, barium, calcium and

copper. The highest critical temperature of ceramics superconductors known till today is 134 K (-139°C), which exhibited by $\text{HgBa}_2\text{Ca}_2\text{Cu}_3\text{O}_8$. Even higher critical temperatures are possible at higher pressure. For bismuth thallium, copper oxides, the critical temperature is 125 K. A few other examples of high temperature superconductors are given in Table (10.4). In almost all high temperature superconductors; the copper oxide is sandwiched between the oxides of the other metals. Copper oxide is common to all the high temperature superconductors, therefore they are also called as Cuprates. The critical temperature 134 K (-139°) is extremely cold by today's standards. However, it can be obtained by using liquid nitrogen (77 K). Liquid nitrogen is cheaper than milk and is readily available. All high temperature superconductors are type II superconductors. Their critical magnetic fields are considerably high. However, BCS theory is not applicable to the high temperature superconductors.

Though the temperature required for high temperature superconductors is easily achievable using liquid nitrogen, some properties of these ceramic superconductors prohibit their applications. These superconductors are brittle, so they cannot be shaped into the wires and coils. Their structures are complex. The properties are highly anisotropic (they change with direction). Another problem with these superconductors is that their critical current densities are low. Thus they cannot carry large currents. Further, necessary crystalline order and purity needs to be developed. They are unstable over longer periods.

It is proposed that the granules of these ceramic superconductors can be arranged in silver tubes. The liquid nitrogen can be circulated around such tubes. These tubes are called as superconducting pipes.

If room temperature superconductor is discovered, then the immediate application that it will have is transmission of the electricity. As I^2R losses are zero, the transmission losses will be zero, thus lot of electricity will be saved. However, as on today, room temperature superconductors are inconceivable.

Isotope effect:

Isotope effect was discovered by Maxwell and Reynolds in 1932. According to this effect, the critical temperature of a superconductor (T_c) is inversely proportional to square root of the isotopic mass of the superconductor

$$T_c \propto \frac{1}{\sqrt{M}} \Rightarrow T_c \sqrt{M} = \text{constant} \quad \dots(10.4)$$

For example, the critical temperature of mercury with an isotopic mass of 199.5 amu is 4.185 K, while the critical temperature for the isotopic mass 203.4 amu is 4.146 K. The isotope effect is consistent with BCS theory. As isotopic mass increases, the lattice vibrations decrease and hence the critical temperature decreases. Isotope effect also indicates that electrons, while moving through the superconducting specimen, do not flow independently, but they interact with the lattice.

Example (10.4):

The critical temperature of a superconductor with isotopic mass 200 is 5 K. Calculate the critical temperature of the superconductor when isotopic mass is 196

Solution:

According to Isotope effect, we have

$$T_c \propto \frac{1}{\sqrt{M}}$$

Thus

$$\frac{T_{200}}{T_{196}} = \sqrt{\frac{196}{200}} \Rightarrow \frac{5}{T_{196}} = 0.9899 \Rightarrow T_{196} = 5.051 \text{ K}$$

Specific Heat Capacity: Specific heat capacity is defined as the heat required to raise the temperature of the unit mass of a given substance by a given amount (usually one degree). Consider Fig 8.4. For normal conductors the specific heat linearly decreases with decrease in temperature. However for superconductors, there is a discontinuity at $T = T_c$. Below T_c , the heat capacity exponentially decreases with the decrease in temperature.

Mechanical effects: It is observed that the critical temperature (T_c) and critical magnetic field H_c change with the stress. Further, below the critical temperature (T_c) there is a slight change in the volume of the superconducting specimen.

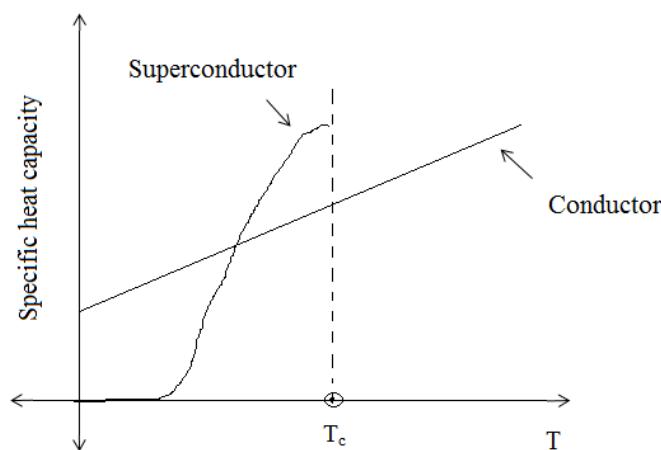


Figure (10.12): Variation of specific heat of conductor and superconductor with temperature

(10.7) BCS THEORY

Ordered flow of Cooper pairs

From the year of discovery (1911) and upto 1957 (46 years) superconductivity remained a mysterious phenomenon. Almost all the results till the date were empirical. However a complete theory of superconductivity was proposed by American Physicists John Bardeen, Leon N Cooper and Robert Schrieffer in a classic research paper published in Physicsal Review Letters in 1957. BCS theory explained all the laws of superconductivity which were known at that time. A few of these were; effect of temperature on supercondcvity (why superconductivity is a low temperature phenomenon), (why superconductivity is destroyed by excess magnetic field), Meissner effect (why magnetic field cannot pass through a superconductor). The BCS theory is based on isotope effect and specific heat of solids.

The key feature of this theory is an ordered state of several conducting electrons in the superconducting lattice. The theory proposes that the electron in a superconducting specimen does not flow without interacting with it. An electron passing through a superconducting specimen attracts the group of positive ions in the lattice towards it. This happens due to Coulombic attraction between the electron and the group of positive ions. Thus the deformed lattice near the electron creates the region of increased positive charge. Thus the lattice near the electron is distorted. Another electron passing nearby the collective group of positive ions, lowers it's energy by getting attracted towards the these positive ions. This also happens due to Coulombic attraction of the distorted lattice of positive ions and the electron passing around.

The lattice vibration which occurs due to an attraction of the electron is called as phonon. This phonon is exchanged by the two electrons which form Cooper pair (Fig 10.13). An electron of wave vector \vec{K} emits a phonon (quantum of lattice energy) which is absorbed by electron of wave vector \vec{K}' .

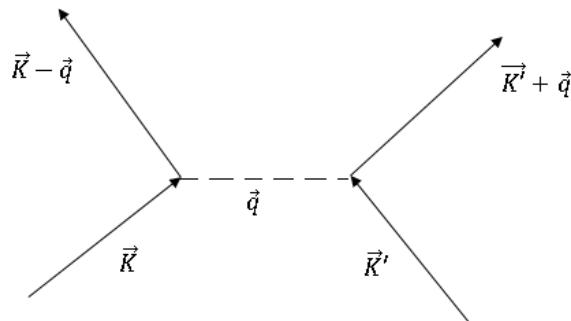


Figure (10.13): Formation of cooper pairs

The electron thus attracted forms a pair with the electron, which attracted the group of positive ions. This is electron-lattice-electron interaction, which results in the formation of pairs of electrons. The pair is called as Cooper pair. It's total energy is less as compared to the total energy of two electrons when they did not form the Cooper pair. The electrons in Cooper pair are negatively charged; hence they experience a repulsive force. But at lower temperature (below critical temperature), the attractive force between the electrons in the Cooper pairs dominates the

repulsive force, as it is energetically favorable. The electrons before the formation of Cooper pairs are free and independent, and they are called as Fermi particles. According to, Fermi Dirac theory and Pauli's exclusion principle, a single energy level can accommodate, only two electrons with opposite spins. Thus the electrons, when they are Fermions, are accommodated in several different energy levels. They have different energies and different quantum states. They are described by different wave-functions. However, when Cooper pairs are formed, they are lowered in the energy. As seen in the Fig. (10.14), Cooper pairs with lowered energy are separated from the Fermi electrons by an energy gap given by

$$2\Delta = E_g = 3.52KT_c \quad \dots(10.5)$$

The energy gap is very small, of the order of 10^{-3} eV, and this is why superconductivity is a low temperature phenomenon. The dependence of energy gap on the temperature T is given by Eqn. (10.6)

$$E_g(T) = E_g(0) \times 1.74 \sqrt{1 - \frac{T}{T_c}} \quad \dots (10.6)$$

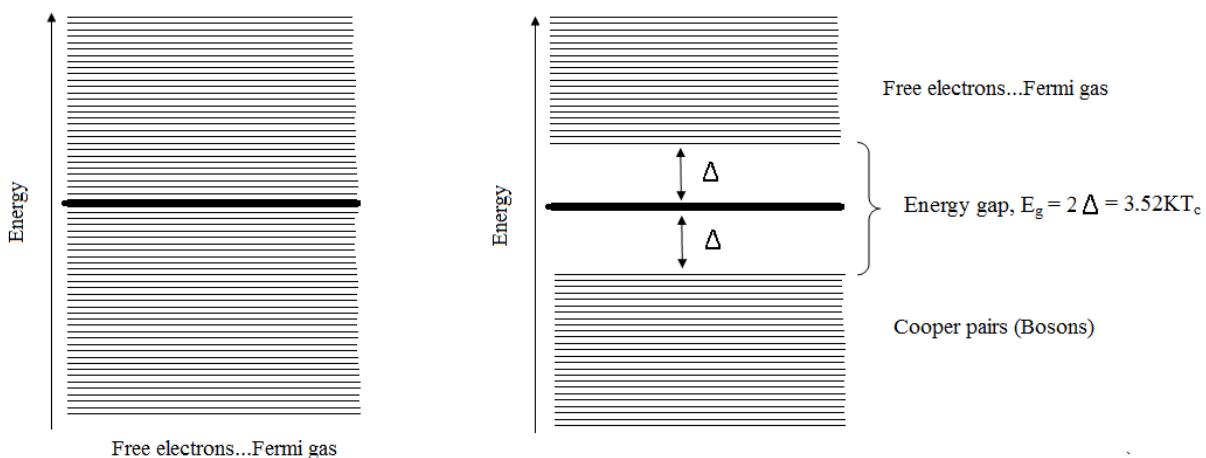


Figure (10.14): Energy level diagram of Fermi gas of electrons and Cooper pairs

As T approaches T_c , the energy gap decreases. When $T \geq T_c$, the energy gap becomes zero and superconductivity is destroyed. At $T = 0$ K, the energy gap is maximum. As temperature increases, some Cooper pairs are broken, the released electrons interact with other Cooper pairs and break them. Thus energy gap decreases as temperature increases.

The electrons (Fermions) can have $-\frac{1}{2}$ or $\frac{1}{2}$ spin. The Cooper pair contains a pair of electrons having antiparallel spins. Thus the net spin of Cooper pair is zero. Hence, unlike free electrons (the Fermi particles), Cooper pair behaves like a Boson. The free Fermi electrons are distributed in several different energy levels. However the Cooper pairs, which are Bosons, can

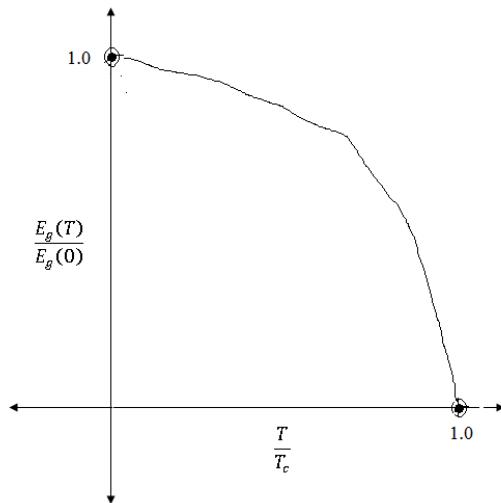


Figure (10.14a): Dependance of energy gap on temperature

be accommodated in a single energy level, no matter what is the number of Cooper pairs. The Pauli's exclusion principle does not apply to Cooper pairs as they are Bosons. In a superconducting lattice, several Cooper pairs of the same quantum state are formed, which occupy the entire region of superconducting lattice. In a volume occupied by one Cooper pair, millions of other Cooper pairs can be accommodated. Since all Cooper pairs are in a single energy state, they can be represented by a single wave function. The group of Cooper pairs distributed over an entire lattice form a giant and ordered system of conducting species. When emf is applied, entire group of Cooper pairs (which have accommodated complete lattice) starts moving. This flow is an ordered flow, where the velocity of all the Cooper pairs is same. Altering the quantum state of single Cooper pair, requires altering the quantum state of all the other Cooper pairs, and hence it is not possible (when $T < T_c$). Thus the group of several Cooper pairs moves through the lattice in a coordinated manner and with the same velocity. The conduction in a superconductor below critical temperature is in ordered state, while for the temperature equal to or above critical temperature it is in disordered state. The equation of current through a superconducting specimen is,

$$I = nev_d A$$

This indicates that, as the density of Cooper pairs is extremely large, even if the velocity of Cooper pairs is small, a substantial current is possible. For the coordinated flow of such slowly moving Cooper pairs, the resistance is almost negligible. The flow of electrons below critical temperature is ordered, while above critical temperature the flow is disordered. This explains, why below critical temperature, the superconductor have zero resistance. Quantum mechanics is the science of subatomic particles, and it mainly works at atomic scale. But, it is observed that for superconductivity, quantum mechanics works at macroscopic level. The entire group of Cooper pairs occupying the area of lattice is represented by single wavefunction. Thus the superconducting state is the ordered state of conducting Cooper pairs. BCS theory thus explains, why superconductivity is a low temperature phenomenon. The energy gap (as given by Eqn 10.5)

is very small. As temperature increases, the energy increases and it breaks the Cooper pairs. BCS theory also explains why superconductivity and magnetic field are natural enemies. The two electrons in any Cooper pair have opposite spins. However when magnetic field is applied and when it reaches the critical value, the electrons are aligned in such a way that their spins become parallel and thus the Cooper pair is broken. Thus the excess magnetic field breaks the Cooper pairs, and superconductivity is destroyed. As we shall see later, in Type II superconductor, the magnetic field through the superconducting specimen increases in the quantum steps of $\frac{h}{2e}$. The presence of the factor '2e' in the formula indicates, that BCS theory (based on Cooper pairs) is consistent with the flux quantization. Isotope effect is also consistent with BCS theory. The critical temperature of the superconducting specimen depends upon the isotopic mass. This indicates that electrons do not flow freely, but they flow by interacting with the superconducting lattice

Flux quantization: We know that in type I superconductor, the magnetic field does not pass through the superconducting specimen. For type II superconductors the magnetic field starts penetrating through the superconducting specimen when H exceeds H_{c1} . It has been observed by A. A. Abrikosov that, the magnetic flux passing through the superconducting specimen is quantized. It increases in the quantum steps given by

$$\phi = n \frac{h}{2e} \quad , n = 1, 2, 3, \dots \text{ (fluxon or fluxoid)} \quad \dots (10.7)$$

The minimum of quantum of flux as given Eqn. (10.7) quantum of flux is called as fluxon or fluxoid. As the magnetic field passing through the superconducting specimen is increased, the number of fluxons increases. In the equation (10.7), n is an integer. The quantity $\frac{h}{2e}$ is the minimum flux quantum that can pass through the superconductor. The magnetic flux through the superconductor increases in the integral (quantum) steps of $\frac{h}{2e}$. After substituting the values of h and e , we get

$$\text{Flux quantum} = \frac{h}{2e} = \frac{6.63 \times 10^{-34}}{2 \times 1.6 \times 10^{-19}} = 2.07 \times 10^{-15} \text{ weber}$$

It is to be noted that flux quantization is significant for type II superconductors. The presence of '2e' in the formula for fluxon confirms the existence of Cooper pairs. The phenomenon of flux quantization was experimentally confirmed by Deaver and Fairbank. In type II superconductors, as the external magnetic field increases, the magnetic field in the superconductor increases in the quantum steps of $\frac{h}{2e}$. The quantum $\frac{h}{2e}$ indicates one fluxon (fluxoid). Thus when the external magnetic field increases, the number of fluxoids/fluxons in the superconductor also increases. These fluxons are surrounded by vortices of supercurrents. Finally when magnetic field exceeds the critical magnetic field H_{c2} , the entire specimen is filled with fluxons.

London Penetration depth: According to Meissner effect, superconductor expels all the magnetic field. Magnetic field inside the superconductor is ideally zero. However it has been proved by F. London and H. London that, the external magnetic field (H) penetrates in the superconductor to a marginal extent. The extent to which the magnetic field penetrates is called

as London penetration depth. Inside the superconductor, the penetrated magnetic field decays exponentially (Fig 10.15).

Consider Eqn (10.8)

$$H(x) = H(0)e^{-\frac{x}{\lambda}} \quad \dots (10.8)$$

(Where $H(0)$ is the magnetic field at $x = 0$)

At $x = \lambda$, we have $H(\lambda) = \frac{H(0)}{e}$. λ is called as London penetration depth (Consider Fig 10.16).

Typically, the London penetration depth is in the range of 3000 Å to 5000 Å. Its exact value depends upon the materials. London penetration depth is a function of temperature. Consider Eqn (10.9)

$$\lambda(T) = \frac{\lambda(0)}{\left[1 - \left(\frac{T}{T_c}\right)^4\right]^{\frac{1}{2}}} \quad \dots (10.9)$$

Where $\lambda(0)$ is the London penetration depth at $T = 0$ K

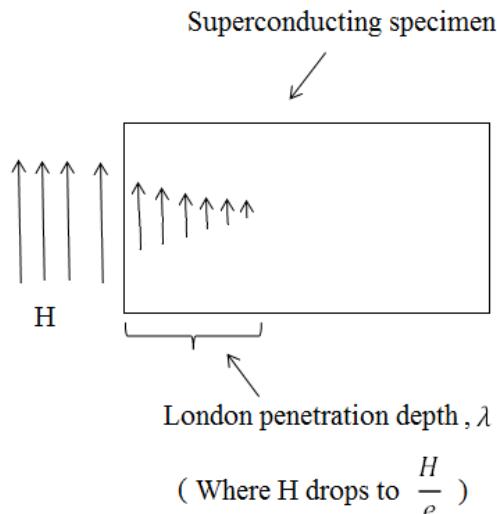


Fig (10.15): Concept of London penetration depth

Eqn (10.9) indicates that at $T = 0$ K, the London penetration depth λ_T has the maximum value given by λ_0 . When T approaches T_c , the ratio $\frac{T}{T_c}$ approaches 1, then $\left(1 - \frac{T}{T_c}\right)$ decreases, and hence the London penetration depth increases. At $T = T_c$, the material is no longer superconducting, the magnetic field penetrates through the entire specimen.

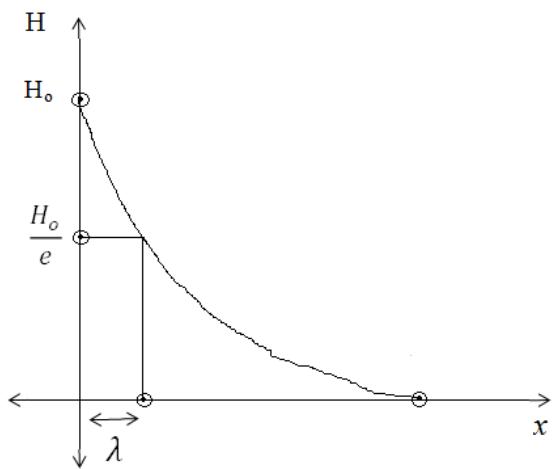


Figure (10.16): Concept of London penetration depth

10.8 APPLICATIONS OF SUPERCONDUCTIVITY:

From ultrathin SQUIDs to giant MAGLEVs

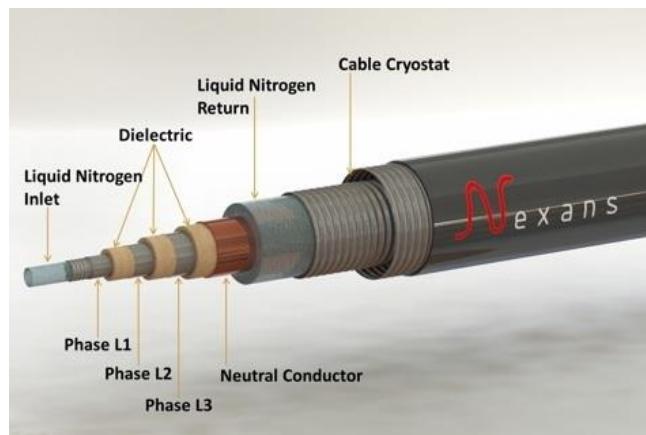
Before 1986, all applications of superconductivity required a lower temperature which could be achieved by liquid helium only. Though high temperature superconductors have been discovered, the materials are brittle, they can not be drawn in to the wires, and they can not carry high currents. In spite of this limitation, a few important applications of superconductivity have been realized.

Superconducting power transmission: In the world $1/5^{\text{th}}$ of the electrical power is dissipated due to I^2R losses. If a room temperature superconductor is realized, then it will be possible to transmit the electrical power without any loss. Further, as there is no resistance, a small emf will be sufficient to maintain the current. High temperature superconductors may replace the conventional conductors in near future

10.10 JOSEPHSON EFFECT

Tunneling of Cooper pairs through a thin insulating barrier

Josephson effect was discovered in 1962 by Brian Josephson when he was the student of Cambridge university (Nobel prize 1975). Josephson effect is exhibited by Josephson junction. Josephson junction consists of two superconductors separated by a thin (1-2 nm) insulating layer (refer Fig 10.17).



Proposed superconducting power transmission cable. It will have zero I^2R losses!

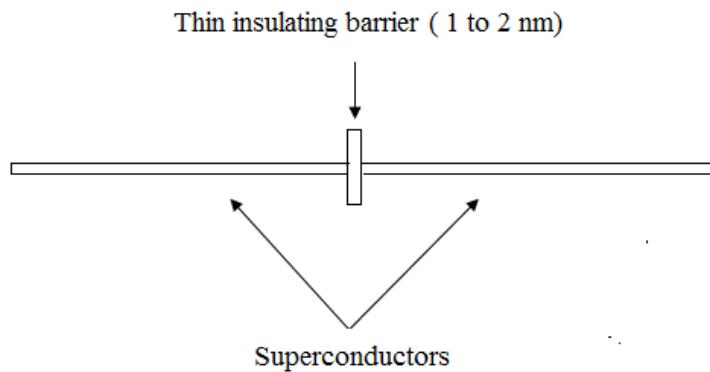


Figure (10.17): Josephson Junction

This superconductor-insulator-superconductor link is also called as weak link. As the insulating barrier is very thin, the wavefunctions of the Cooper pairs on either side overlap and thus the tunneling becomes possible. The Cooper pairs in the superconductors overcome the resistance of the barrier due to tunnel effect. The tunneling of Cooper pairs through the Josephson junction is called as Josephson effect. There are two types of Josephson effects, one is DC Josephson effect (DC current when no voltage applied) and AC Josephson effect(AC current when DC voltage is applied).

The mathematics is optional

The DC Josephson effect: The Cooper pairs flow across the junction even if no voltage is applied. The tunneling results in to a dc current. Due to thin insulating barrier, there occurs a phase difference between the wavefunctions of Cooper pairs on the either side of the junction. The dc current is given by

$$I_s = I_c \sin \phi \quad \dots(10.10)$$

Where ϕ is the phase difference between the wavefunctions on either side of the junction and I_c is the critical current when the voltage is zero. I_c depends upon thickness and width of the insulating layer and temperature.

AC Josephson effect: When DC voltage is applied, there occurs an additional phase difference of $\Delta\phi$ between the wavefunctions of the Cooper pairs on the either side of the insulating barrier. This results in to an oscillating (AC) current of high frequency through the junction. This current is given by

$$I = I_c \sin(\phi + \Delta\phi)$$

It can be shown that when a DC voltage V is applied across the junction, the energies of the Cooper pairs on the either side differ by $2eV$. Then

$$\Delta\phi = 2\pi t \left(\frac{2eV}{h} \right)$$

Substituting $\Delta\phi$ in the formula for I , we get

$$I = I_c \sin \left[\phi + 2\pi t \left(\frac{2eV}{h} \right) \right]$$

I is the AC current flowing due to a DC voltage V . The frequency of this AC current is given by

$$\nu = \frac{2eV}{h} \quad \dots(10.11)$$

(we have $h\nu = 2eV$)

Thus a photon of frequency ν is emitted by the junction, when a DC voltage is applied. The frequency of this radiation is very high

Example (10.5).

A DC voltage of $100 \mu V$ is applied across the Josephson junction. Calculate the frequency of radiation thus emitted

Solution: We have

$$\nu = \frac{2eV}{h}$$

$$\nu = \frac{2 \times 1.6 \times 10^{-19} \times 100 \times 10^{-6}}{6.63 \times 10^{-34}} = 4.83 \times 10^{10} \text{ Hz}$$

The eqn (10.11) indicates that, the fundamental constant $\frac{e}{h}$ can be calculated by using Josephson effect. Thus Josephson effect provides a method to determine the fundamental constant $\frac{e}{h}$. Josephson junction also forms the basis of SQUIDs (Superconducting Quantum Interference Devices), which have many applications.

Superconducting Quantum Interfering Devices (SQUIDs): SQUID is basically an ultrasensitive magnetometer. It used to detect and measure extremely weak magnetic fields which can be as small as 10^{-15} T. Such weak magnetic fields are generated by human brain and heart. The SQUID is essentially a superconducting ring having combination of one or two Josephson junctions(Refer Fig 10.18). There are two types of SQUIDs, AC SQUID and DC SQUID. DC SQUID consists of two Josephson junctions. There occurs an interference of the currents passing through the junctions. The entire device is kept in liquid helium.

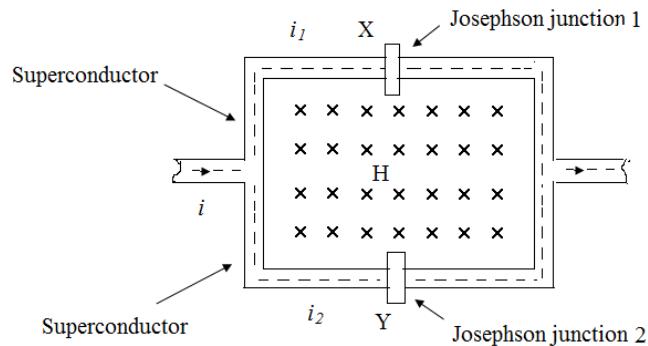
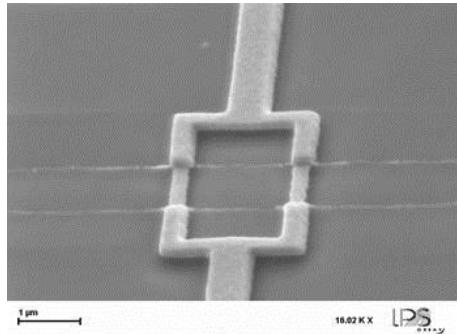


Figure (10.18): SQUID

As shown in the Fig. (10.18), the current i passing through a superconductor is divided into two currents i_1 and i_2 . These currents pass through the Josephson junction 1 and 2. The two Josephson junctions have different thickness. In absence of any magnetic field through the loop, there occurs no phase difference between the supercurrents (Cooper pairs). However, when the magnetic field is made to pass through the loop, there occurs a phase difference between the currents coming from the two Josephson junctions. The net current is thus the function of the magnetic field passing through the loop. In practice, the voltage across the loop is measured instead of the current. SQUID is thus magnetic flux to voltage transducer. The smallest magnetic field that can be measured by the SQUID is 10^{-15} T. Such weak magnetic fields are generated by human brain and heart. The analysis of brain by using SQUID is called as magnetoencephalography (MEG). SQUIDs also find applications in geology and nondestructive testing. A few more applications of SQUIDs are

- a. Computer memories
- b. Processing elements in digital computers
- c. Geomagnetic abnormalities (earthquake prediction)
- d. Sensitive voltmeter...can measure the voltage, as small as 10^{-16} V
- e. Detector of infrared radiation.

- f. SQUIDs are used as logic gates (switches) with switching time of picoseconds (on the contrary, the switching time is nanoseconds)

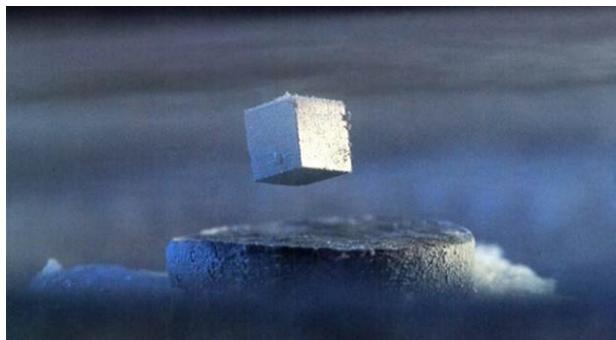


An ultrathin SQUID



SQIUD being used for monitoring Brain activity

MAGLEV: Maglev means Magnetically Levitated Train. The speeds of ordinary trains are limited due to friction between the wheels and the railway track. The friction also results in heat. Both these limitations can be overcome by allowing the Maglev to float on the premagnetized railway track. Current is passed through the railway track. At the bottom of the compartments, there are superconducting magnets. Due to Meissner effect, there occurs a repulsion between the railway track and the compartment. The train is thus lifted up (approximately 4 inch). As friction is avoided, the MAGLEV runs efficiently, smoothly and with greater speeds (500 km/hr at present). MAGLEVs are now competing with air traffic.



Demonstration of the Meissner effect. The Superconductors are perfectly diamagnetic and it repels the magnet



The Shanghai Maglev Train, also known as the Transrapid, is the fastest commercial train currently in operation and has a top speed of 501 km/h.

Electromagnets, Transformers, Generators, Solenoids, Motors (large scale applications).

The devices which are used to generate heavy magnetic fields can employ superconductors instead of conventional ferromagnetic materials. Stronger magnetic fields can be generated without eddy current losses and hysteresis losses. Therefore, such devices can be lighter, smaller and efficient. The devices based on conventional materials are cumbersome, and they demand large electric power. The heavy magnetic fields generated by superconducting magnets can be used in several applications such as MAGLEV, magnetic fusion, accelerators and Magnetic

Resonance Imaging (MRI). The accelerators at Fermi lab and LHC (Large Hadron Collioder) make the use of superconducting electromagnets. As discussed earlier, all these applications employ type II superconductors.

A few more applications of superconductors are

- Meissner effect form the basis of frictionless bearings
- Superconducting computer is being perceived
- There is a limitation for making the ICs small (I^2R losses). With superconductors, it will be possible to make the IC extremely small and compact. Cramming/crowding more circuits per unit area. This low temperature electronics is called cryotronics



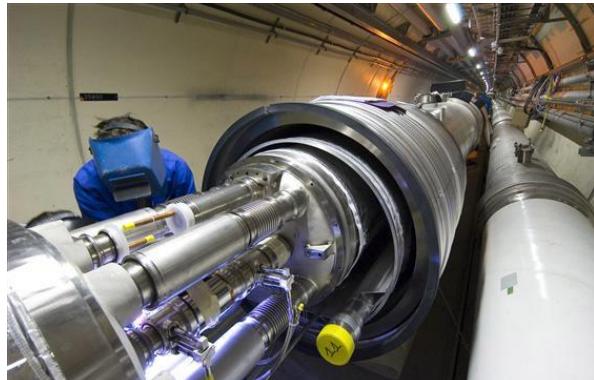
Superconducting electromagnets produce considerably stronger magnetic field than ordinary electromagnets



Large Hadron Collider (LHC). The largest machine in the world, detected Higg's Boson



Superconducting electromagnets are used in MRI, a Nobel prize winning diagnostic technique



Superconducting electromagnets are used in Large Hadron Collider (LHC)

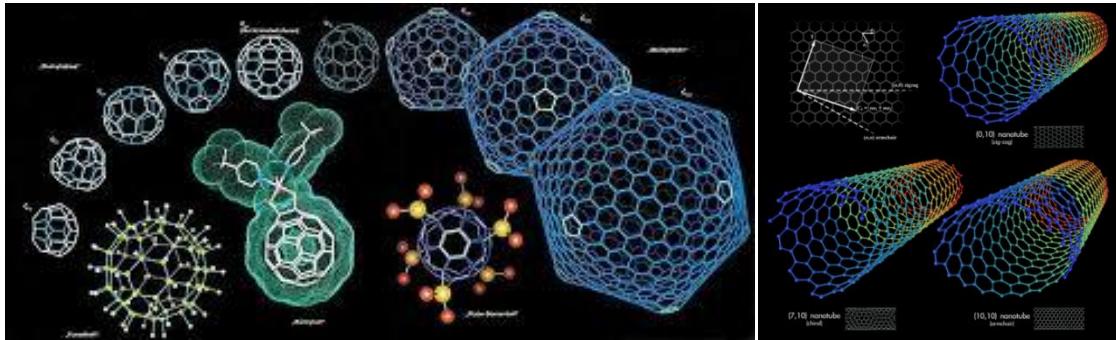
REFERENCE BOOKS

1. Fundamentals of Physics by Halliday, Resnick, extended, Wey
2. Concepts of Modern Physics, Arthur Beiser, Tata McGraw Hill
3. A Textbook of Engineering Physics, Avadhanulu and Kshirsagar, S. Chand, 10th Edition
4. Solid State Physics, N. W. Ashcroft and N. D. Mermin, (CBS Publishing Asia Ltd.)
5. Introduction to Solid State Physics, Charles Kittel, (John Wiley and Sons.)
6. Introductory Solid State Physics, H. P. Myers, (Viva Books Pvt. Ltd.)
7. Solid State Physics, H. Ibach and H. Luth, (springer-Verlag).
8. Fundamentals of Solid State Physics, J. R. Christman, (John Wiley and Sons.)
9. Solid State Physics, A. J. Dekkar, (Prentice Hall). 7. Physics of Semiconductor Devices, S. M. Sze (John Wiley and Sons.)
10. Introduction to Superconductivity: Second Edition, Michael Tinkham, Courier Corporation,
11. The Physics of Superconductors: Introduction to Fundamentals and Applications, V.V. Schmidt, Springer Science & Business Media
12. Superconductivity: An introduction, Philippe Mangin, Rémi Kahn, Springer, Technology & Engineering

WORLD WIDE WEB

1. www.superconductors.org/
2. www.superconductivity.eu/
3. <http://www.superconductorweek.com/>

Physics of Nanoparticles



The photograph on the left shows buckminsterfullerene (Buckyball, C_{60}) and that on the right shows carbon nanotubes (CNTs). Buckminsterfullerene is a spherical fullerene molecule with the formula C_{60} . Its structure resembles a soccer ball having diameter in the range of nanometer. Buckyballs are harder than diamond and they find applications in solar cells, in human bodies as antioxidants and also in drug deliveries. Carbon nanotube is a tube shaped material made of carbon, having diameter measuring nanometer scale. The thinnest carbon nanotubes have a diameter of 3 \AA , while the longest CNTs have length of 18.5 cm. CNTs show a unique combination of superior mechanical, thermal and electrical properties. CNTs can be used for power and data transmission, batteries, solar cells and hydrogen storage etc. Both buckyball and carbon nanotubes are nanostructures and known to exhibit fantastic properties and applications. What are nanostructures? how they are synthesized? and what are their applications?

The answer to these questions is in this chapter

Index

11.1 NANOTECHNOLOGY: INTRODUCTION

“There is plenty of room at the bottom”...Richard Feynman (1959)

11.2 NANOMATERIALS:

Quantum mechanics plays a decisive role at nanometric scale

11.3 PROPERTIES OF NANOMATERIALS

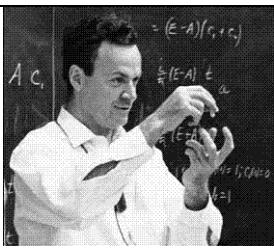
Mystic properties just due to decrease in dimension

11.4 SYNTHESIS OF NANOPARTICLES

Top down and bottom up approaches

11.5 APPLICATIONS OF NANOPARTICLES

Materials, devices and machines with superior properties



Richard P. Feynman: He studied at Massachusetts Institute of Technology and Princeton. After receiving his Ph.D. in 1942, he helped develop the atomic bomb at Los Alamos, along with many other young Physicists. When the war over, he first went to Cornell and, in 1951, to the California Institute of Technology. The Genius made several dominating contributions to Fundamental Physics, and Technology. A few of these include Quantum electrodynamics, the relativistic quantum theory, Feynman diagrams, explanation of the behavior of liquid helium near absolute zero, quantum computing, elementary particle theory etc. When requested to identify the causes of Challenger accident, his opinion differed from NASA engineers, however, he held his view, until he was proven correct. During his lecture given in American Physical Society, he suggested...‘*There's Plenty of Room at the Bottom*; And this became the beginning of the Nanotechnology. For his contributions in development of Quantum electrodynamics he shared Nobel prize in 1965. His three volumes Lectures on Physics has stimulated and enlightened both students and teachers since its publication in 1963.

11.1 NANOTECHNOLOGY: AN INTRODUCTION

“There is plenty of room at the bottom”...Richard Feynman (1959)

Now a days we are all familiarized with the word nanotechnology. The word ‘nano’ refers to small, may it be a nanocar or may it be a washing machine using nanoparticles or may it be a housing project. The word ‘nano’ is greek in origin and it means dwarf/small. In science, nano means billionth of a meter (10^{-9} meter). In last few years we have seen many advantages of making the things small. A few years back the computers were too big and they required buildings for their installation. The radios were also big and required large power. But in last few years we have seen a revolution in technology, due to which we have come across many compact but versatile gadgets such as calculators, small but efficient personel computers, laptops, compact and smart mobile phones, CD players, pocket transistors etc. Though nature consists of only 118 elements in the periodic table, several compounds, alloys and composites have facilitated our life. In our modern life we use many nature-made and man-made smart materials. The materials play a dominant role in wonderful electronic systems, communication tools, transport vehicles, textiles, utensils, agricultural materials, medicines etc. The progress of a human race requires newer, smart and novel materials. It has been now well established that one way to obtain materials with fascinating properties is to make them smaller than 100 nm. It has been observed that when at least one dimension of a material is made smaller than 100 nm, most of its properties, including mechanical, optical, electrical, chemical, magnetic etc., change to a significant extent. Further, below 100 nm, the properties are size and shape dependent. For example, in 1857, Michael Faraday showed that gold, which appears yellow in bulk (macro) form, shows red color in nano form. It has also been observed that CdS nanoparticles appear red when their size is 10 nm, however, at 4 nm, 3 nm and 2 nm, they show orange, yellow and white color respectively. Titanium oxide show self-cleaning effect, Zinc oxide shows UV blocking properties and clay particles become hard when reduced to nanometric scale. Thus it is possible to tailor the properties of the same material by changing its size in nano-regime. The nanodevices

or nanomaterials can be made small by two approaches, one is top-bottom approach, and another is bottom to top approach. The first approach is mastered by human beings (for ex. the integrated circuits are made smaller and smaller) while the bottom-top approach is already mastered by nature. In top-bottom approach, the devices and materials are carved and cut to make them smaller and smaller till they acquire nano form. In bottom-top approach the devices and materials are assembled by bringing the individual atoms together. In his famous visionary and prophetic lecture entitled '*There is plenty of room at the bottom*' delivered on 29th December 1959 in front of American Physical Society, a renowned Physicist and Nobel Laureate, Richard Feynman suggested mimicking nature's bottom-top approach to make nano-metric materials, devices and machines. The term nanotechnology was coined by Norio Taniguchi in 1975. *Nanotechnology deals with the design, production, characterization and application of structures, devices, and systems by controlling shape and size at the nanometer scale*

Nanotechnology in the past: Michael Faraday was the first Physicist who produced stable gold nanoparticles in 1857 and showed that, while gold appears yellow in color in its macro-form, when brought in to nano-form, it shows magenta red color. The red solution containing gold nanoparticles is still preserved in the British Museum in UK. Similarly in mediaeval ages the decorative glass windows with beautiful designs in old churches and palaces indeed used nanoparticles of iron, cobalt, nickel, gold, silver etc. In 9th Century silver and copper nanoparticles were used in Mesopotamia for generating glittering effect on the surface of the pots. Old photography is an example of nanoparticles (18th and 19th centuries). Photographic film is an emulsion, a thin layer of gelatin containing silver bromide. Light decomposes silver bromide producing nanoparticles of silver which are the pixels of the image. However, nanotechnology is considered as a modern technology, as in the past, powerful microscopes did not exist and a systematic theory explaining the properties at nanometric scale was not available.

Why Nanotechnology? One of the greatest significance of nanotechnology is that the behavior of a material can be engineered at nanometric scale. Further the properties of the materials at the macro scale considerably change when they are doped with nano-materials. Nanotechnology is still in its infancy stage but has lot many potential applications. Principle domains which will be affected by development of nanotechnology are

- i. **Materials:** New lighter but stronger and harder materials, which are more durable and resistant, lighter and less expensive
- ii. **Electronics:** Electronic components with smaller and smaller size, which may result in powerful electronic gadgets and faster switches in computers
- iii. **Energy:** Vast increase in solar energy generation. There is a considerable concern now on developing renewable, inexpensive, clean or pollution-free energy sources and solar cells based on nanomaterial open up new challenge for the scientists and technologists.
- iv. **Health and Biotechnology:** Prevention of diseases, diagnosis and treatment. Drug delivery to specific organs.

A few industries which will be progressed due to nanotechnology are pharmaceuticals, cosmetics, consumer appliances, textile, information technology, communications, security and safety, space exploration etc. It is said that nanotechnology will bring next industrial revolution.

Nanotechnology offers better products with exacting specifications, better built, long lasting, cleaner, safer and smarter products for the homes, for communications, for medicine, for transportation, for agriculture and for industry in general. Nanotechnology is a highly multidisciplinary field which is related with several other disciplines such as Physics, Chemistry, Biology, Material science and all Engineering fields. This chapter aims at study of nanomaterials, their properties, methods of their synthesis, and their applications.

10.2 NANOMATERIALS:

Quantum mechanics plays a decisive role at nanometric scale

There are fundamentally two reasons behind exhibition of characteristically different properties by materials when they are brought to nanometric scale. The first is significant increase in surface to volume ratio when the materials are made small. We know that surface area is proportional to r^2 , while the volume is proportional to r^3 . Thus surface to volume ratio is proportional to $\frac{1}{r}$. Thus when a material or a device is made smaller, its surface to volume ratio increases. The drastic increase in surface to volume ratio makes a material more functional, thereby causing a change in its physical and chemical properties. Second reason exists in quantum mechanics. According to quantum mechanics, the properties of a subatomic particle trapped in a potential well (a typical 'box' or a nanocluster of atoms) significantly change with the size of the potential well or cluster. This is well explained by a formula for energy of motion of a particle in potential well

$$E_n = \frac{n^2 h^2}{8mL^2} \quad \dots(11.1)$$

The equation 11.1 clearly indicates that as the width (L) of the potential well changes, the energy level diagram and hence the properties of particle change. We know that classical mechanics fails at nanometric scale. Nanotechnology is thus governed by laws of quantum mechanics. Nanoparticles of nanodevices have their one (or all) of the dimensions within a range of 1nm to 100 nm. To get a rough idea about a nanometric scale, dimensions of the some of the objects are given in the table (11.1)



Michel Faraday was the first to synthesize Gold nanoparticles in 1857. These are still preserved in British Museum in UK. The particles are still stable.

We know that 1 gm of a material contains approximately 10^{23} atoms. The dimensions of the nano-materials are so small they contain very few atoms. For ex. 1 nm contains 10 hydrogen atoms stacked in a line. Before we proceed further, it is necessary to understand the difference between a nanoparticle and a nano-cluster. The size of cluster is usually around or less than 1 nm and it contains less than 1000 atoms. The size of nanoparticle is between 1 to 100 nm and it typically contains 1000 to 10^6 atoms.

Sr. No.	Object	Typical dimension
1	Diameter of Sun	1,393,000 km
2	Diameter of an earth	128,000 km
3	Height of Himalaya mountain	8,848 m
4	Height of a man	1.65 m
5	Fly	1 cm
6	Single human hair	80000 nm
7	Red blood corpuscles	10,000 nm
8	Limit of eye's ability to see	10,000 nm
9	E coil bacteria	2000 nm
10	Visible spectrum	700 to 400 nm
11	Virus	20-250 nm
12	Size of a nanoparticle	1-100 nm
13	Quantum dot	5 nm
14	DNA	2 nm
15	Carbon nanotube	1.3 nm
16	Buckyball	1 nm
17	Size of hydrogen atom	0.1 nm

Table (11.1): Some objects and their typical dimensions

11.3 PROPERTIES OF NANOMATERIALS:

Mystic properties just due to decrease in dimension

Nanomaterials are the fragments of bulk materials. When the size of the material is decreased below a critical size (typically 100 nm), the properties change significantly. The critical size depends upon the structure and properties of the material. Below the critical size, the properties not only become size dependent but also shape dependent. Thus it is possible to tailor the properties of nanomaterials. Herewith, we will review mechanical, electrical, optical, structural and magnetic properties of nanomaterials.

Optical properties:

Metallic nanoparticles

It is now a well-established fact that the color of the material changes when it is brought in to nanometric form. This phenomenon was used by medieval workers 2000 years back to make

tinted glasses. Such glasses are found to be used in old churches, palaces and houses. Various beautiful colors such as red, pink, blue, green were imparted to the transparent glasses by dissolving small amount of (< 5%) of various particles such as those of gold, silver, cobalt, nickel etc. Of course the physics behind this first and oldest use of nanotechnology was not known at that time, the attempts of systematic understanding of the relation of colors with size of particle began with Michael Faraday, who produced the nanoparticles of gold in 1857. He reduced chloroauric acid (HAuCl_4) using citric acid [$\text{CH}_2(\text{COO})_2\text{H}_2\text{O}$] and showed that the nanoparticles of gold thus produced intense red color in nano form as compared to yellow color in bulk form. It is also known that zinc oxide has superior UV blocking properties as compared to its bulk form. Therefore nano zinc oxide is used in sunscreen lotions.

A satisfactory explanation of change in color due to change in size was given by G. Mie in 1908. His explanation is based on Maxwell's electromagnetic theory. According to him, when a beam of light of intensity I_o and wavelength λ passes through a medium embedded with uniformly distributed particles, the intensity reduces due to scattering of light. This intensity is given by

$$I = I_o e^{-\mu x} \quad \dots(11.2)$$

Where μ is the extinction coefficient which depends upon number of particles in the medium, volume of the colloidal particles and extinction cross section of a particle. μ is given by

$$\mu = \frac{N}{VC_{ext}} \quad \dots(11.3)$$

Where N is the number of particles in the medium, V is the volume of colloidal particles. It can be shown that C_{ext} and hence μ depends upon R , the radius of particle embedded in the medium. It can also be shown that presence of particles in the medium gives strong resonance band in the visible region.

Another explanation for change in color due to change in particle size exists in quantum mechanics. The bulk materials are characterized by energy bands. When they are converted into nano form, the energy band diagram is transformed to discrete energy levels. Using quantum mechanics (Eqn 11.1), it can be shown that the spacing between the energy levels changes with the size of the nanoparticle. Thus the transitions and hence the color changes with the size of nanoparticle.

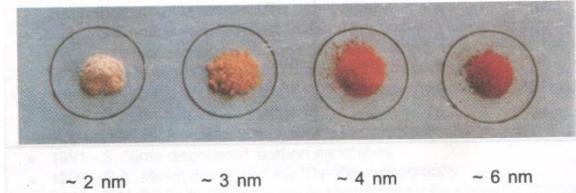
Semiconductor nanoparticles: It is well known that bulk semiconductors are characterized by typical energy gaps. A few examples are given below

Semiconductor	$E_g(\text{eV})$ for bulk
CdS	2.58
GaAs	1.52
InP	1.42
InSb	0.24

It has been shown that the energy gap of the semiconductor increases when it is reduced to nanoform. Further, it is also observed that the energy gap increases with the size of semiconductor nanoparticle. An example of GaAs nanoparticles is given below

Size of GaAs nanoparticles	20 nm	10 nm	5 nm	2 nm
Energy gap in eV	1.42	1.46	1.61	2.78

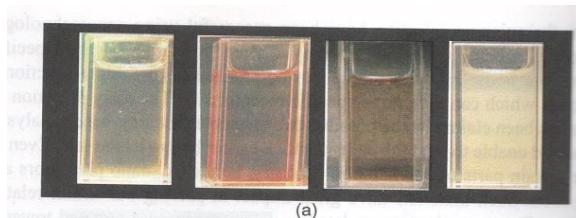
When the energy of an incident photon coincides with the energy gap of a semiconductor, it is absorbed. In such a case the UV-Visible absorption spectrum shows an absorption peak at corresponding photon wavelength. As energy gap increases with the size of nanoparticle, the absorption peak in the UV Visible spectrum shows a blue shift in the UV-Visible spectrum. It has been shown that Cd_3P_2 is a dark brown semiconductor with an energy gap of 0.5 eV. As it is



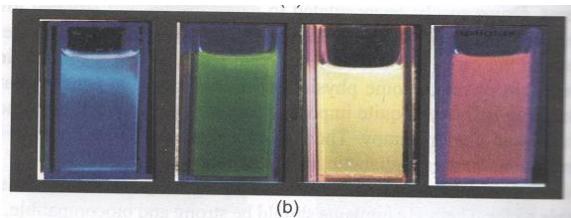
CdS nanoparticles: color changes with size of nanoparticles



Gold nanoparticles: color changes with size of nanoparticles



- a. The solution containing CdSe particles is under normal day light.
- b. The fluorescence produced by CdSe nanoparticles as the size of particles reduces (red to blue) can be seen. These are same solutions as in (a) but are illuminated with hand held UV lamp





Quantum dots tuned between ultraviolet and infrared on the electromagnetic spectrum. Quantum dots are tiny particles of a semiconductor material (usually based on cadmium or zinc) ranging from 2 to 10 nanometres (about 50 atoms) across. Owing to their size they have very different properties compared with larger pieces of the same material, as predicted by quantum theory.

converted in to nanoform, the color changes from brown to red to yellow and finally white with particle size changing from $\sim 30 \text{ \AA}^\circ$ to 15 \AA° . For 15 \AA° particles the band gap increases to 4 eV. This is also true for CdS nanoparticles. In bulk form the energy gap of CdS is around 2.58 eV and it is orange in color. As the CdS nanoparticles become smaller, the energy gap increases, it becomes yellowish and ultimately white

Luminescent properties of nanoparticles

Some materials including nanomaterials when excited with external source of stimulus like electron, light etc, they emit light in the visible, UV and IR range. This is called as luminescence. Many nanoparticles exhibit enhanced luminescence as compared to their bulk counterparts. Some materials like silicon which are not luminescent in their bulk form become luminescent in nanoform. As band gap depends upon the size of the nanoparticle, the luminescence can also be tuned to desired wavelength in nanomaterials. This property is used in display channels. Some materials become luminescent when nanoparticles are doped in them.

Electrical properties:

Electrical properties of nanoparticles (such as quantum wells, quantum dots, quantum wires/carbon nanotubes etc.) are different than their bulk counterparts. The conductivity of a material depends upon number of electrons, effective mass, defects as well as scattering etc. Conductivity is given by

$$\sigma = \frac{Ne^2\tau}{m} \quad \dots(11.4)$$

Where σ = conductivity of an electron, N = number of electrons, e = electronic charge, τ = relaxation time (time between collision of electron with two ions) and m = mass of an electron

The resistivity of a material (metal/semiconductor) can be measured by using ohm's law. The resistivity of metals are low ($10^{-6} \Omega - cm$), for semiconductors resistivity are medium (few $\Omega - cm$), while the resistivity of insulators are high ($> 10^3 \Omega - cm$). We know that typically the graph of current (I) Vs voltage (V) for conductors; semiconductors and insulators is a straight line (Fig. 11.1). However, if the material is reduced to a nanomaterial (size below 100 nm), then there is a deviation from this behavior. The I-V characteristics of a quantum dot (a material whose all three dimensions are reduced to nanometric scale) is not a straight line but it appears like a staircase. It can also be observed from Fig (11.1 b) that, from zero up to certain low bias voltage, there is no current. This phenomenon is called as *Coulomb Blockade*. This is due to the fact that unless a voltage of $e^2/2C$ is applied, a single electron cannot tunnel through the

quantum dot. The graph appears like a stair case due to repeated tunneling of electrons.

At this stage, it is essential to understand the concept of quantum well, quantum wire and quantum dot. Out of three dimensions of a material, if one dimension is reduced to nanometric scale, then electron's motion is confined to one dimension, however, electron is free to move in other two directions. Such system is called as a quantum well. The energy of an electron is quantized along a direction along which it is confined. Consider a wire which has no restriction over a length but whose diameter is reduced to a nanometric scale. In such a case the motion of

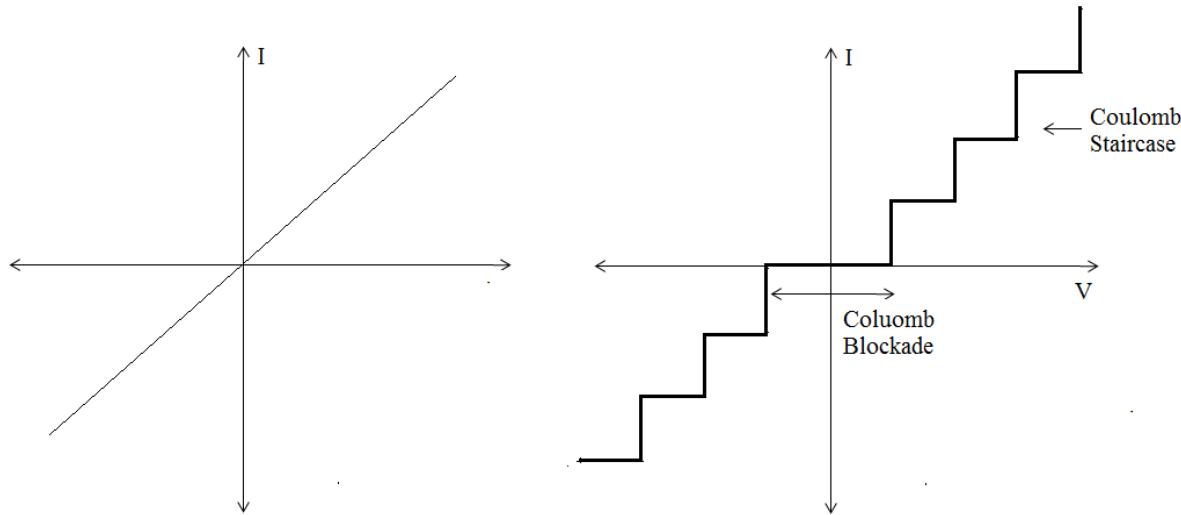


Figure 11.1 I-V Characteristics of (a) conductor (b) quantum dot

electron is confined to two transverse directions and the energy is quantized. Such wires are called as quantum wires. Due to quantization, the classical formula of resistivity given by $\rho = R \frac{A}{l}$ is not applicable to quantum wires. The electrical conductance of quantum wires is quantized in the multiples of $2e^2/h$ where e is the electronic charge and h is the Planck's

constant. Carbon nanotube (CNT) is an example of quantum wire. If all three dimensions of a material are reduced to a nanometric scale, it is called as a quantum dot. The phenomenon of Coulomb Blockade and Coulomb Staircase is observed frequently in case of quantum dots and

nanoparticles. The properties of quantum dots are intermediate to bulk materials and discrete molecules. The electronic properties of quantum dots are also closely related to their size and shape

In general, the resistivity of nanocrystalline materials is greater than their bulk analogs having microcrystalline boundaries. This is because electrons will suffer scattering to a greater extent when the crystalline boundaries are smaller in size and more in number. It is also a fact that the resistivity of a multicrystalline material is larger than a single crystal material. It is observed that nanomaterials show superconducting properties under certain conditions. A '*single electron transistor*' having size less than 10 nm has been demonstrated. Its I-V characteristics is different than that of a conventional transistor. Such single electron transistors have good potential for low power high density integrated circuit applications.

Magnetic properties:

Magnetic properties of a material are due to spin and orbital motion of the electrons around the nucleus. Corresponding to these motions there are magnetic moments, which are vectorially added. Depending upon the electronic configuration of the atoms, interaction between the atoms and the resulting magnetic behavior, the materials are categorized to five classes namely diamagnetic, paramagnetic, ferromagnetic, antiferromagnetic and ferrimagnetic. Magnetic materials are important in technology due to their diversified applications such as electronic circuits, transformers, motors, sensors, information storage, memory devices and medical field. We have seen that in semiconductor electronics, the integrated circuits are becoming more compact and the electronic devices are becoming smaller. It is observed that in case of magnetic materials like Fe, Co, Ni, Fe_3O_4 etc, an interesting magnetic behavior is exhibited when these materials are reduced in size below a critical size (typically 100 nm)

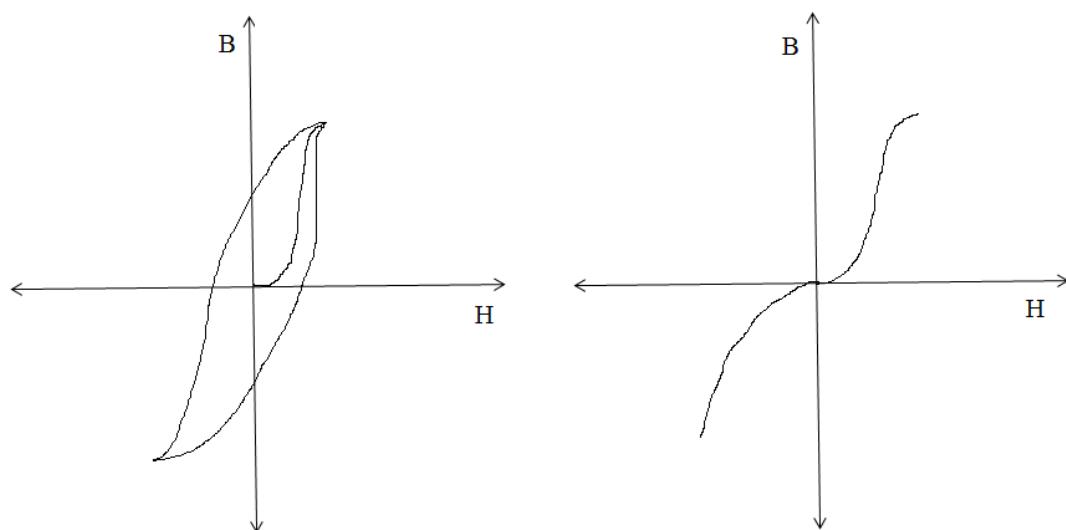


Figure 11.2 (a) The hysteresis loop of a ferromagnetic material (b) B-H curve of a nanomagnetic material

All ferromagnetic materials are characterized by a typical property called hysteresis (Fig 11.2 a). The B-H curve of these materials displays a typical loop called as hysteresis loop. The ferromagnetic materials are characterized by spontaneously magnetized regions called as domains. Such domains are formed for the minimization of magnetostatic energy. In a given domain the magnetic moments of all the atoms are aligned and thus the domain has a net magnetic moment. However, in a given material, the magnetic moments of different domains are in different directions and thus the net magnetic moment of the material is smaller than that would have been if all the domains were aligned. When placed in magnetic field, the magnetic moments are aligned and the material is said to be magnetized. During demagnetization, the material does not retrace the same path as it was during magnetization, and thus a hysteresis behavior is exhibited. A negative force called as coercive force needs to be applied to demagnetize the material completely. When the size of ferromagnetic material is reduced below its critical size (typically 100 nm), the multi-domain formation is not energetically favored and the nanomaterial behaves like a single domain. Thus there are no multiple domains, domain-walls and boundaries. Therefore nano-ferromagnetic material does not display hysteresis behavior. There is no coercive force. The typical B-H curve of a nano-ferromagnetic material is as shown in Fig 11.2 b. Such materials are called as superparamagnetic materials.

It is known that, when the material is reduced to nanometric form, its surface to volume ratio increases significantly. At the surface there is breaking of symmetry. The lattice constant, coordination number on the surface are different for nanomaterial. Therefore a material which is not ferromagnetic in bulk form shows ferromagnetism in nanoform. For nanomagnetic materials it is observed that the saturation magnetization increases with decrease in particle size. Nanomagnetic materials exist in different forms such as multilayers, spin valves, magnetic tunnel junctions (MTJ) and the oxide materials.

Multilayers: Multilayers are the layers of materials of nanometric thickness in which several layers of two materials are alternatively deposited on each other. Multilayers of several materials including semiconductors, metals, insulators and organic materials are possible. The magnetic multilayers have a typical property. In magnetic multilayers the layers of ferromagnetic or antiferromagnetic materials and metals are alternatively deposited. It has been demonstrated by two Nobel laureates, Albert Fert and Peter Grunberg (Nobel prize in Physics in 1996) that magnetic multilayers exhibit Giant Mageto Resistance (GMR). This means that the resistance of magnetic multilayers changes significantly (by 50 to 60%) when magnetic field is applied to them. A typical example of magnetic multilayer is Co-Cu (alternative nanolayers of Co and Cu). The property of GMR is invariably used in today's computers for data storage.

Based on magnetic multilayers and GMR, there are two other nano devices...one is called as **Spin Valve** and another is called as **Magnetic Tunnel Junction (MTJ)**. Spin valve is a tri-layer of magnetically soft (easy to magnetize) material, copper and magnetically hard (difficult to magnetize) material. Spin valves are used in computer read heads. The ability of spin valves to detect small magnetic fields has led to an increase in the data storage capacity of memory devices. Magnetic Tunnel Junction (MTJ) is made up of at least two magnetic layers separated by insulating tunneling barrier. MTJ exhibit high magneto resistance. If metal in a magnetic multilayer is replaced by metal oxide, then the change in resistance on application of magnetic field is even larger than that of magnetic multilayers. Such magnetoresistance is called as Colossal Magnetic Resistance (CMR).

Structural properties:

We know that different materials have different crystalline structures such as cubical, hexagonal, rhombohedral, tetragonal etc. One of the most common techniques used to determine the crystal structure of a material is X-ray diffraction. The lattice constants can also be determined using X-ray diffraction data. Interestingly, the structure of a material changes when it is brought to nanoform. The lattice constant also changes. Thus nanomaterials are not just the fragments of bulk materials but they have different structures also. For example, it is known that ZnS in bulk form has sphalerite (cubic) structure; however, the nanoparticles of ZnS having a size of 1.4 nm show a structural disorder like a liquid. It is observed that the structure of nanoparticles changes on changing the temperature and pressure. It is also observed that the pressure required to change the structure of a nanoparticle is considerably larger than that required to change the structure of their bulk counterpart.

Mechanical properties:

The mechanical properties of materials include stress, strain, Young's modulus, hardness, ductility etc. The mechanical properties of the material depend upon the strength of the bonds (such as ionic metallic or covalent). The mechanical properties also depend upon impurities (such as C, O, N, P, S etc) and the defects, grain boundaries and dislocations. When the material is reduced to nanometric form, it behaves like a single crystal. The properties such as Young's modulus, density, hardness change when the material is reduced to nanometric form.

It has been observed that for magnesium, in polycrystalline form (grain size $> 1 \mu\text{m}$), the Young's modulus is 4100 N/mm^2 , however, when it is reduced to nanometric scale (grain size $\sim 12 \text{ nm}$), the Young's modulus decreases to 3900 N/mm^2 . Further, the Palladium in polycrystalline form, the Young's modulus is $12,300 \text{ N/mm}^2$, but when it is reduced to nanometric form the Young's modulus is decreased to 8800 N/mm^2 . Refer table 11.1

Material	Density (kg/m³)	Young's modulus (10⁹ N/m²)
Steel	7860	200
Diamond	3510	1035
Carbon Nanotubes (CNT)	2600	1280

It appears that the Carbon Nanotubes (CNT) are the strongest materials on the earth.

The density of nanocrystalline pellet is often low due to some pores left when powders are compressed to form pellets. However, if the sintering is done at high temperature, then the density of nanocrystalline pellets approaches the density in polycrystalline form.

The dependence of hardness on the grain size of the material also changes when the material is reduced to nanometric form. Refer Fig 11.3 a and b. It has been observed in case of copper and palladium that in microcrystalline form, the hardness increases with the grain size, however, at nanoscale the hardness increases on decreasing the grain size.

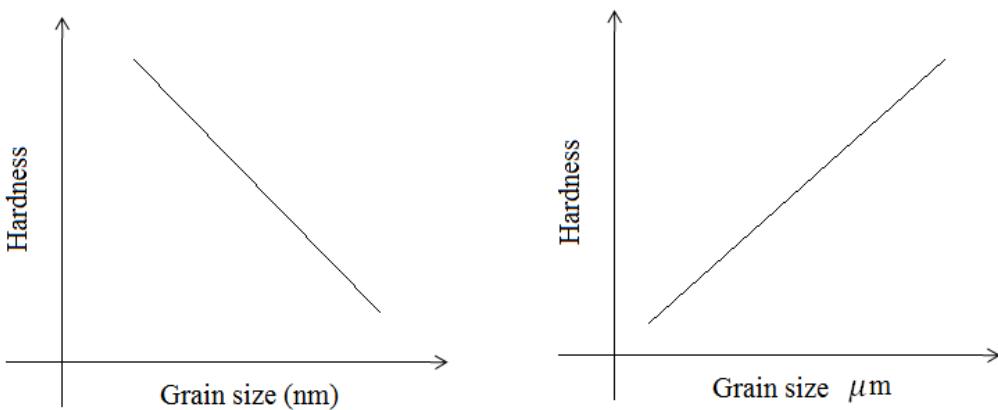


Figure (11.3) Dependence of hardness on grain size

10.4 Synthesis of nanoparticles

Top down and bottom up approaches

Nanoparticles exist in variety of forms such as colloids, clusters, powders, tubes, rods, wires, thin films etc. Depending upon the requirement, such as material of interest, size, shape, quality and quantity, there exist variety of methods to synthesize the nanoparticles. As nanotechnology is an interdisciplinary science; physical, chemical, biological and hybrid methods are used to synthesize the nanoparticles. Broadly, there are two approaches to synthesize the nanoparticles. One is top down approach in which the bulk material is successively cut to bring it to nanometric form. Another approach is bottom up approach where individual atoms are manipulated to form a nanoparticle. The list of methods of synthesis of nanoparticles continues to grow. Here, we discuss a few methods to synthesize the nanoparticles.

Physical methods:

Top down approach

Mechanical method:

This is one of the simplest methods to make nanoparticles of metals and alloys in the form of powder. Using this method the nanoparticles of metals and alloys are made. This is top down approach. Different types of mills such as planetary, vibratory, rod, tumbler etc. are used. More than one container are filled with the hardened balls of steel or tungsten carbide and the bulk material (whose nanoparticles are to be made) in the form of powder or flakes (having size less than 50 μm). Initially the material particles are arbitrary in size and shape. The lid of the container is made tight. The size of the container to be chosen depends upon the quantity of the nanoparticles to be produced. The ratio of mass of balls to mass of powder is 2:1. The cylinder

is less than half filled. If it is filled to more than half, the efficiency of milling process decreases. The containers are rotated with high speed around their own axis (spin motion). The spinning cylinders are also rotated around certain axis like a planetary motion. This is just like the motion of planets which revolve around themselves and also around the Sun. Due to such motion this mill is also called as *planetary ball mill*. Due to planetary motion, the powder is forced to towards the walls and pressed against the walls, because of centrifugal force. But due to spinning motion the powder is moved to other region of the cylinder. If the size of the balls is increased then the impact energy increases due to which size of nanoparticles is decreased however, defects are also introduced. The impurities may be introduced due to balls and the presence of air or gases. If impurities due to gases are to be avoided, highly pure gases should be used. Due to the milling process, the temperature may be increased to 100 to 1100° C. To reduce the temperature, liquid nitrogen is used. By controlling the speed and duration of rotation, the nanoparticles with size of few nm to few 100 nm are produced. The process can produce nanoparticles of few milligrams to few kilograms within a time of few minutes to few hours. Using this method, nanoparticles of Co, Cr, W, Al-Fe, Ag-Fe, Ni-Ti can be synthesized.

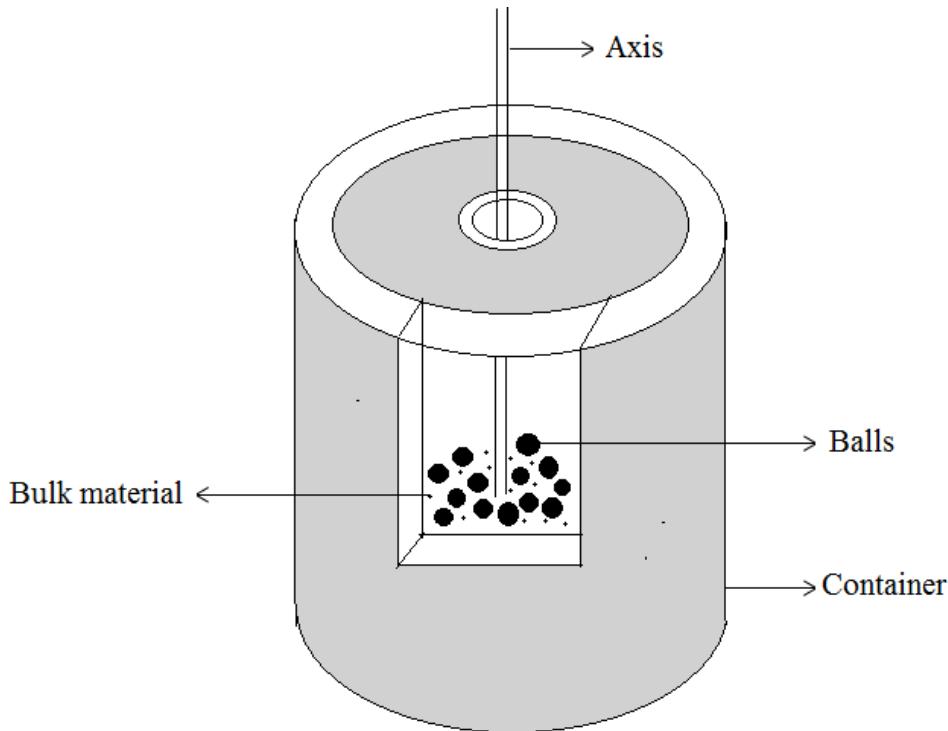


Figure (11.4): Planetary ball mill

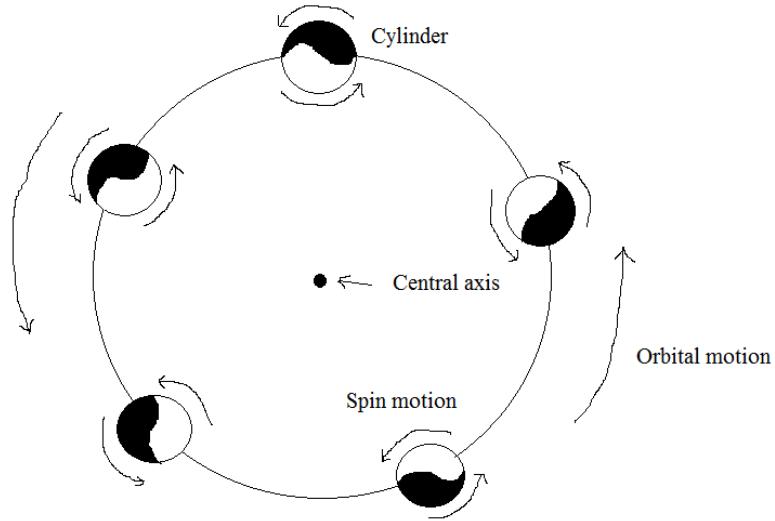


Figure (11.5): Spin and orbital motions of planetary ball mill

Electric Arc Deposition(OPTIONAL)

The important and special forms of nanoparticles such as fullerenes, Carbon nanotubes can be synthesized at mass scale by using this method. The method is simple and commercially viable.

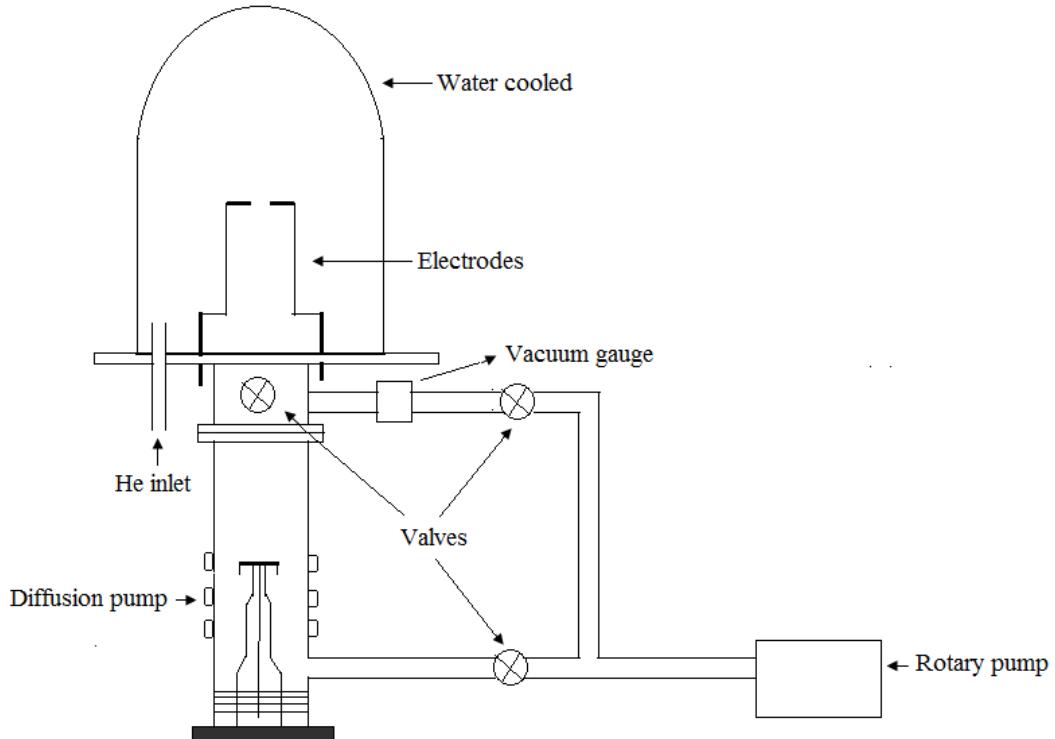


Figure (11.6): Set up for electric arc deposition

Refer Fig (11.6). Following are the essential requirements of electric arc deposition method

- i. Water cooling
- ii. Vacuum (using rotary pump and diffusion pump)
- iii. Provision of inert gas such as helium or reactive gas

The heart of this system is a pair of electrodes (anode and cathode) having a gap of approximately 1 mm between them. Anode acts as a source of nanomaterial (for ex. graphite). High current of approximately 50 to 100 amperes is passed through the electrodes using a low voltage power supply (~ 12-15 volts).

The main part of this method involves striking an arc between the electrodes. The arc strikes when the current is passed. Due to this, the electrode burns and the material coated on the anode evaporates. The evaporated material is deposited on cathode as well as in the inner part of the chamber. Due to evaporation of the material on the anode, the gap between the electrodes increases. This is unfavorable for striking the arc between them. A system is therefore necessary to adjust the gap between the electrodes. During the discharge, the temperature reaches to ~ 3500 °C, therefore cooling system based on water is necessary. The material used for the synthesis of nanoparticles exists on the anode; however, if catalysts are to be used, a separate system for evaporation is necessary.

This method has been successfully used for the synthesis of fullerenes in large quantity. In this case, both the electrodes are made of graphite. The formation of fullerenes or Carbon Nano Tubes (CNTs) depends upon the helium pressure. Fullerenes are formed at comparatively low helium pressure while for Carbon Nano Tubes, high helium pressure is necessary. The fullerenes are obtained by purification of the soot deposited on the inner walls of the chamber. They are not found on the cathode, whereas Carbon Nano Tubes are formed only on the central part of the cathode and not on the walls of the chamber. Formation of the carbon nanoparticles is also possible. However, this method is particularly suitable for the syntheses of fullerenes and Carbon Nano Tubes.

Evaporation techniques can also be used to synthesize the nanoparticles. Mainly, there are two techniques based on evaporation; one is Physical Vapor Deposition (PVD) and the second is Chemical Vapor Deposition (CVD). Both are discussed below.

Physical Vapor Deposition (PVD):(OPTIONAL)

This technique is based on the evaporation of the materials. Refer Fig (11.7). The PVD set up consists of following parts

- i. Material to be evaporated
- ii. Crucible for heating the material
- iii. A system to evaporate the material (heating or sputtering)
- iv. Inert gas or reactive gas for collision with the material vapor
- v. A cold finger (liquid nitrogen cooled) on which clusters of nanoparticles can condense
- vi. A scrap to scrape out the nanoparticles deposited on the cold finger
- vii. Piston-anvil (an arrangement for compacting the nanoparticle powder and formation of pellets)

viii. Vacuum system (purity of the nanoparticles depends upon the vacuum)

The materials required to synthesize the nanoparticles can be metals or metal oxides. These are heated or sublimated in crucibles. The material in the crucible can be vaporized with resistive heating or electron beam sputtering. Resistive heating is accomplished by winding a high resistance wire around the crucible and passing high current through it. In electron beam sputtering, high energy electron beam is directed on the material. During evaporation the density of the evaporated material near the crucible is high and the particle size is small (<5 nm). Such particles prefer to be in low energy state and thus an interaction of these particles with each other may result in increase in the size of the particles. Therefore these particles need to be taken away from the source and blown towards the cold finger, where they condense. This can be done by forcing the inert gas towards the source. This gas takes away the nanoparticles from the area in the vicinity of crucibles. The size and size distribution of the nanoparticles is governed by three factors which are, the rate of evaporation, pressure of the gases, and the distance between the crucibles and the cold finger. Lesser the distance between the crucibles and cold finger, smaller

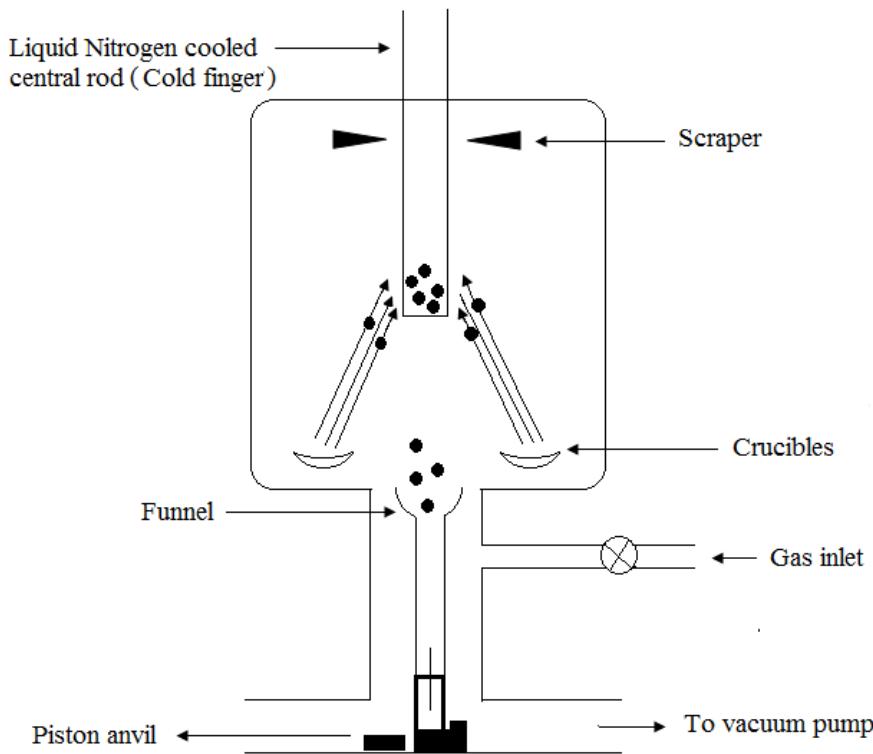


Fig (11.7): Schematic diagram of the Physical Vapor Deposition (PVD) technique

is the size of the nanoparticles. The growth of nanoparticles also occurs during their movement from crucibles to the cold finger. The nanoparticles condensed on the cold finger are scraped using scrapers. The scraped nanoparticles then fall in the funnel, from where they are then taken to the piston-anvil arrangement where they are compacted to form pellets. If reactive gases like

O_2 , N_2 , H_2 , NH_3 are used then compounds of metals and the reactive gases are formed. Nanoparticles in large quantities are obtained by repeating the evaporation and deposition process.

Chemical Vapor Deposition (CVD):

Chemical Vapor Deposition involves the evaporation and deposition of materials on the hot substrate and their chemical reaction with the substrate. Under certain reaction conditions, this results in the formation of nanoparticles (generally in the form of thin films). The nanoparticles of metals or metallic compounds can be formed. The reactant in the form of vapor are pumped in the reaction chamber by using carrier gas, then vapor or gas is transported towards the substrate maintained at high temperature (usually ~ 300 to $1200^{\circ}C$). The gas or vapor is deposited on the substrate, where it undergoes a chemical reaction at appropriate sites. The chemical reaction also results in to the byproducts which have to be suitably removed from the substrate. The quality of the product is governed by gas pressure and substrate temperature. CVD technique involves relatively simple instrumentation, ease of processing and economic viability.

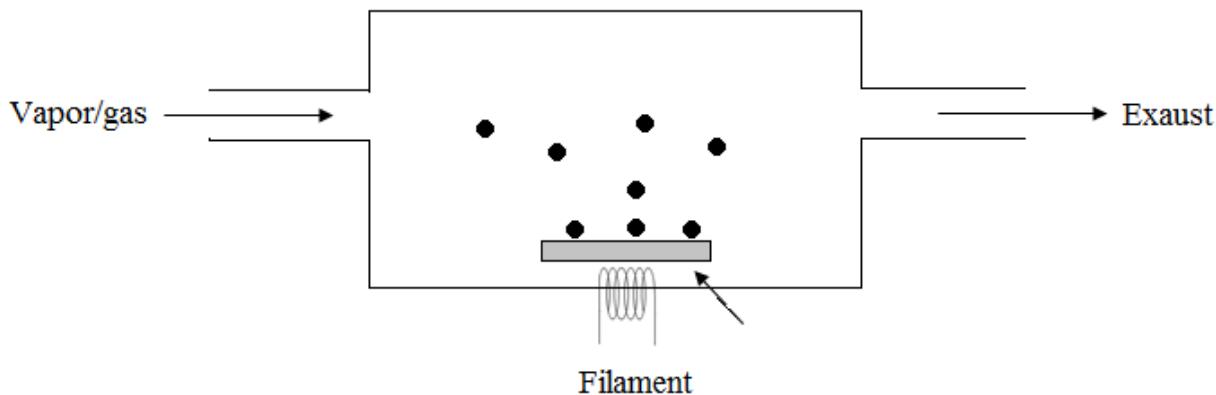


Figure (11.8): Chemical Vapor Deposition (CVD)

Laser Pyrolysis: Pyrolysis means decomposition at high temperature. Laser pyrolysis or laser-assisted decomposition is based on the decomposition of gases using laser. It is possible to decompose the gases such as C_2H_2 , C_2H_4 , $Fe(CO)_5$, using high power laser (such as cw CO_2 laser). The decomposed atoms interact with each other, grow in size to acquire nanoform and then get deposited on silicon substrate. This method is particularly useful for synthesis of Carbon Nano Tubes, however, the nanoparticles of Al_2O_3 , Si_3N_4 can also be synthesized using this technique. The characteristics of nanoparticles such as size and size distribution depend upon the gas pressure and substrate temperature. Presence of inert gases such as helium or argon is necessary. Refer Fig (11.9), which shows the block diagram of laser pyrolysis set up. The apparatus also involves the pressure and temperature control units. Nanoparticles in the form of thin films are obtained by this method.

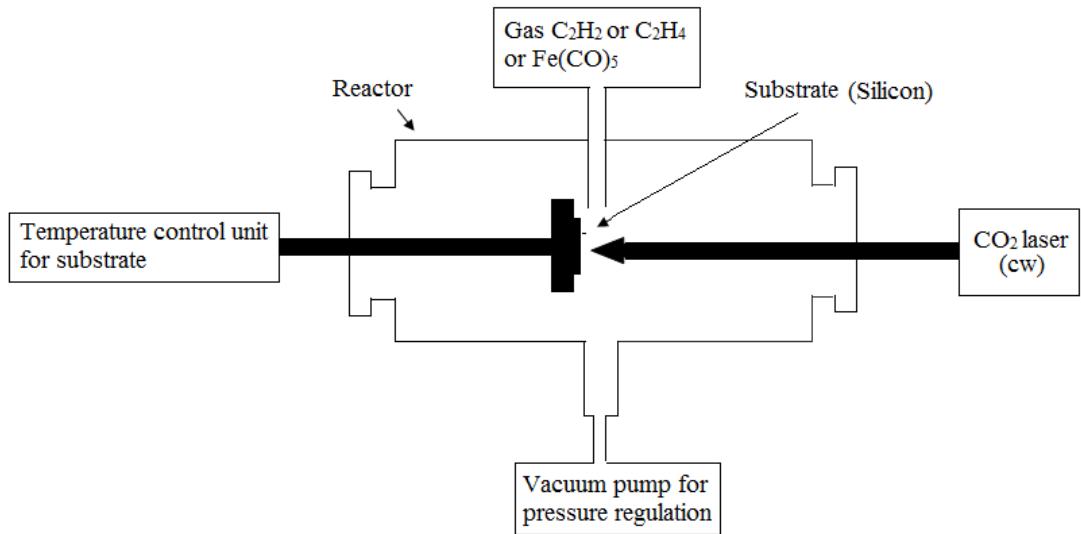


Figure (11.9): Laser Pyrolysis

Chemical Method: (OPTIONAL)

Bottom up approach

Chemical methods have several advantages over the other methods. Some of these are enlisted below

- i. Simple
- ii. Inexpensive, as less number of instruments are required
- iii. Low temperature synthesis is possible
- iv. Doping of foreign atom or ion is possible during synthesis
- v. Nanomaterial can be obtained in large quantities
- vi. Varieties of sizes and shapes are possible
- vii. Very narrow size distributed nanoparticles are possible
- viii. The nanomaterial is obtained in the form of liquid, but can be easily converted in to powder or thin films

One of the chemical methods for synthesizing nanoparticles is through colloidal route. The nanoparticles are obtained in the form of colloids. The nanoparticles are obtained in solution and then centrifuged, filtered and dried and thus obtained in powder form. Colloidal dispersions in organic liquids are called as organosols and dispersions in water are called as hydrosols. Chemical reactions are used to synthesize the nanoparticles. Depending upon the chemicals and reaction conditions nanoparticles of different shapes and sizes can be made

What are colloids?

We know that matter exists in three phases: solid, liquid and gas. Colloids are the class of materials in which two or more phases of the same or different materials coexist with the dimensions of at least one phase less than micrometer. Colloids can have various shapes such

that sphere, fiber, tubes, rods or films etc. Nanoparticles in colloidal form are a special case of colloids, where at least one dimension of the colloid is of the order of nanometers. There are several examples of colloids

- i. Liquid in gas: Fog
- ii. Liquid in liquid: fat droplets in milk
- iii. Solid in liquid: toothpaste
- iv. Solid in solid: Tinted glass
- v. Gas in liquid: Foam
- vi. Water or oil in the porous rocks
- vii. Bio-colloids
 - a. Blood: Corpuscles in serum
 - b. Bones: Calcium phosphate embedded in collagen

Colloidal particles are generally charged. The particles are kept in suspension by the repulsive forces between them. Colloids of metals, semiconductors and insulators, organic and inorganic species are possible. Synthesis of nanoparticles by colloidal route dates back to 19th century, when Michael Faraday synthesized gold nanoparticles. These nanoparticles are still stable and kept in British Museum, London

Synthesis of Colloids: Colloids are the particles suspended in some host matrix. Using this method, metal, semiconductor, insulator and alloy particles of various shapes and sizes can be obtained in aqueous and non-aqueous media. The nanocolloids are stabilized either due to Columbian repulsion or surface passivating molecules. This method is also called as wet chemical method.

For synthesis of nanoparticles by colloidal route, a tri-necked glass reactor is necessary (Refer Fig 11.10. it has a provision to insert the thermometer and pH meter , and reactants/precursors and inlet and outlet for the inert gases like argon or nitrogen. The necessity of inert gases is for avoiding the oxidation during the reaction. There is also a provision for removing the products at suitable stages during the reaction. The flask is kept on a magnetic stirrer with temperature control. The reactants are stirred using the magnetic needle.

Growth of nanoparticles (LaMer Diagram): The synthesis of nanoparticles by chemical route is a complex process. It is quite necessary to control the nucleation, growth and saturation of nanoparticles so that monodispersed nanoparticles (of nearly same size) can be obtained. This can be understood with the help of a LaMer diagram (refer Fig. 11.11).The size of nanoparticles is principally governed by concentration. When we increase the concentration of reactants, at certain stage, the concentration C_0 is reached. At this stage the formation of nuclei begins. There is no precipitate at this concentration. The concentration is further increased up to C_N , above which there is super-saturation. At C_N the rate of formation of nuclei is maximum. When nuclei formation reduces, again, C_0 the minimum concentration for

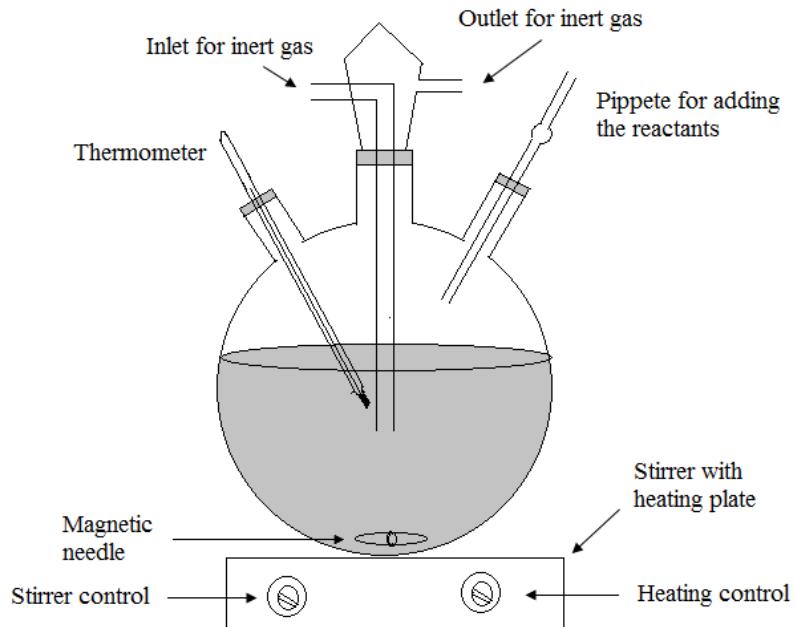


Figure (11.10): Experimental set up for colloidal route method

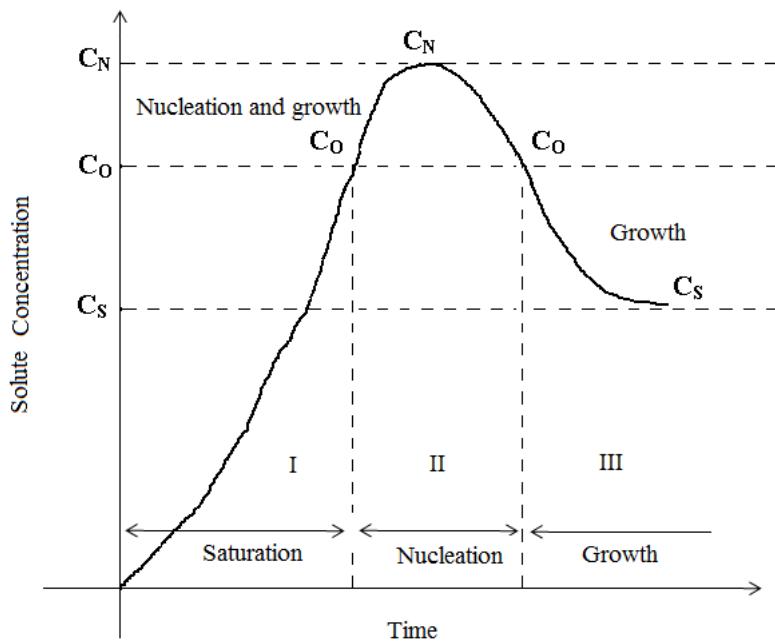


Figure (11.11): LaMer Diagram

nuclei formation is reached. No new nuclei are formed at this stage and then the concentration C_s is reached. At this concentration (C_s), equilibrium is reached. If new nuclei are formed during the growth of nanoparticles then the particles with large size distribution are obtained. To avoid this, the concentration needs to be controlled in such a way that no fresh nuclei are formed after

C_N . If this is not avoided, then different nuclei will be at different stages of growth and nanoparticles of large size distribution are formed.

Larger nanoparticles grow at the cost of smaller particles, as the larger particles are more stable than the smaller ones. This is called as *Ostwald Ripening*. The cause of formation of the larger particles is reduction of the surface free energy. In addition to growth of particles to bigger size, it has been experimentally found that smaller particles come together to form aggregates. This is called aggregation. Aggregation also reduces the surface free energy.

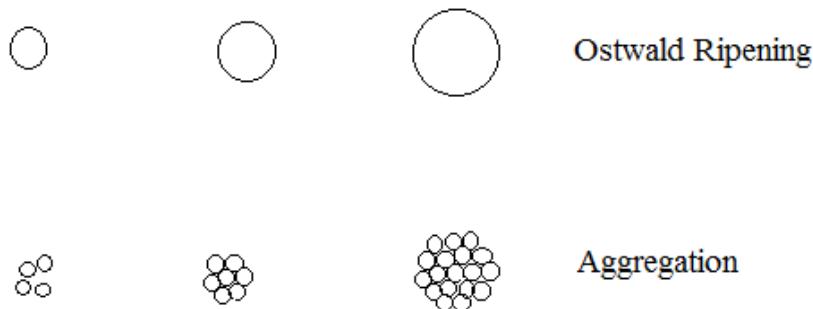


Fig (11.12): Growth and aggregation of colloidal nanoparticles.

Ostwald ripening and aggregation proceed side by side. (Refer Fig 11.12)

Synthesis of metal nano particles by colloidal route: Colloidal metal nanoparticles can be synthesized by reduction of the metal salt or acid. For ex. highly stable gold nanoparticles are obtained reduction of chloroauric acid (HAuCl_4) by trisodium citrate ($\text{Na}_3\text{C}_6\text{H}_5\text{O}_7$). The reaction takes place in a following manner.

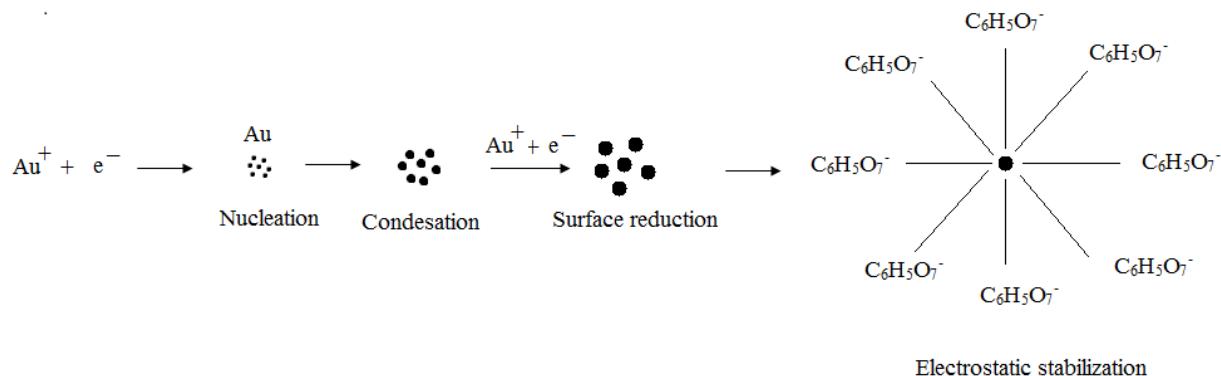


Figure (11.13): Stabilization of gold nanoparticles

As discussed earlier, the steps involved in the synthesis are nucleation, condensation and Capping. The stabilization of the gold nanoparticles takes place due to oppositely charged citrate ions. The reaction is carried out in water using a set up as shown in Fig (11.10). As gold

nanoparticles are formed, the color changes to intense red, magenta etc. depending upon their size. The stabilization of the gold nanoparticles takes place due to repulsive Coulombic interaction. Thiol or other capping molecules are also used for capping the gold nanoparticles.

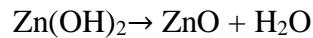
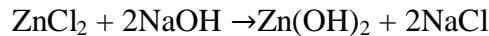
This method can also be used to synthesize the nanoparticles of some other metals such as silver, copper or palladium. Nanoparticles of required size and shape and size distribution can be obtained by using appropriate precursors and by controlling the reaction parameters such as concentration, temperature, pH, duration of reaction etc. Bimetallic or alloy nanoparticles can also be obtained by such methods.

Synthesis of semiconductor nanoparticles by colloidal route: In addition to elemental semiconductor such as silicon and germanium, there are compound semiconductors. Examples are, gallium nitride (GaN), gallium arsenide (GaAs), indium arsenide (InAs), zinc sulfide (ZnS), silicon carbide (SiC), indium selenide (InSe), cadmium sulfide (CdS) etc. The nanoparticles of such compound semiconductors can be obtained by using colloidal wet chemical method. For this purpose, salts of these compound semiconductors are required. The synthesis of nanoparticles takes place due to co-precipitation method.

Reaction of Zinc Chloride ($ZnCl_2$) with Na_2S or H_2S (gas) results in the formation of nanoparticles of ZnS.



The nanoparticles of ZnO can be synthesized by using following reaction routes.



Passivation is necessary to prevent coagulation or ripening of colloids. As Coulomb interaction is not sufficient to stop this process, stearic hindrance needs to be used for the stabilization of the nanoparticles. The nanoparticles obtained in this way can be converted in powder form (drying) or thin film.

Several other methods are used to synthesize the nanoparticles. These include

- i. Melt mixing
- ii. Ionized cluster beam deposition
- iii. Laser vaporization (ablation)
- iv. Sputter deposition
- v. DC sputtering
- vi. RF sputtering
- vii. Magnetron sputtering
- viii. ECR plasma deposition
- ix. Ion implantation
- x. Molecular beam epitaxy
- xi. Langmuir-Blodgett method

- xii. Microemulsions
- xiii. Sol-Gel method
- xiv. Hydrothermal synthesis
- xv. Sonochemical synthesis
- xvi. Microwave synthesis
- xvii. Synthesis using microorganisms
- xviii. Synthesis using plant extracts
- xix. Synthesis using DNA

10.4 Applications of nanoparticles

Materials, devices and machines with superior properties

We know that the properties of materials and devices change at nano-level. Thus it is possible to make materials, devices and machines with improved properties using nanotechnology. Nanomaterials possess unique physicochemical properties. As a result, nanotechnology finds applications which span almost every sector of science and technology. The fields in which nanotechnology finds applications are

- i. Automobiles
- ii. Energy
- iii. Electronics
- iv. Medical
- v. Domestic appliances
- vi. Space and defense
- vii. Cosmetics
- viii. Textile
- ix. Sports and toys

Herewith, we will discuss some notable applications of nanotechnology. The list of applications discussed here is not exhaustive, as newer applications of nanotechnology are being discovered almost daily.

a. Electronics:

Electronics has become an indispensable component of modern technology oriented life. Wristwatches, calculators, portable computers, laptops, mobile and even satellites, space missions and internet are all based on electronics and related areas.

In last century, electronics has passed through some important stages which have led to miniaturization, speed and reduction of cost. First was invention of transistor, which replaced bulky, power consuming and costly vacuum tubes. Second was invention of integrated circuit (IC) and third is VLSI (very large scale integration). According to Moor's law, the electronic devices are shrinking day by day. After 2000, the electronic devices and materials have become

so tiny that they have entered in the realm of nanotechnology. Due to this, now the properties of such nano-electronic devices have become both size and shape dependent.

We know that electron has two properties; one is charge and second is the spin. The electronics so far is based on the charge properties of electron. An entirely new form of electronics, named spintronics (or magnetoelectronics) is coming up. Spintronics is based on the spin of the electron. The electronic devices based on the charge properties can be manipulated using electric field. However, the spintronic devices are manipulated using magnetic field. The materials used for spintronic devices are the semiconductors doped with the magnetic materials (such as Fe, Co and Ni). The examples of the spintronic devices are Single Electron Transistor, Spin LEDs, Spin Valves, Spin FETs, and Magnetic Tunnel Junctions, Giant Magneto Resistance (GMR), Colossal Magneto Resistance (CMR), Q-bits for Quantum Computer etc. These spintronic devices are fast, compact and relatively cheap as compared to electronic devices. The advantage of using the spin property of electrons instead of their charge is that, unlike the charge the spin is not easily destroyed by scattering from collisions with impurities and defects

Spin valves are already being used in the computer memories. This has increased data storage capacity of the computers. Further, in the spin based computers, the memory is nonvolatile in the sense that, the data is not lost even though there is a power failure. The screens of the flat panel TV and computer are based on nanoparticles. A new type of display called as electro-chromic display, based on the nanoparticles of tungsten oxide has been realized.

b. Energy:

Energy is the primary requirement of the global population. Due to increase in world population, the energy demands are also increasing. At present the energy is being harnessed from coal, oil and natural gas and it is being used in transportation, communication, agriculture, industry, and houses etc. The energy is also being harnessed from dams and nuclear reactors. The materials being used for conventional energy are depleting. Further, the conventional techniques of energy production have some serious disadvantages such as pollution and global warming. Search for newer ways of harnessing the energy by nonconventional methods are thus necessary. An intense research in this area is going on. Nanotechnology shows hopes in this area also. Basically, three areas of energy generation are based on nanotechnology; one is energy from hydrogen, second is solar cells and third is the rechargeable batteries.

- i. **Hydrogen as a fuel:** Hydrogen is an effective source of energy and can be used as a fuel. It can be used as a fuel in automobiles. Attempts are being made to harness hydrogen by decomposing water (H_2O) from the sea water using sunlight. In this process, nanomaterials are being used as photo catalysts. Although hydrogen is an effective source of fuel, one of its disadvantages is that it is highly combustible and it catches the fire easily. Thus storage of the hydrogen is difficult. The Carbon Nano Tubes (CNTs) will provide an effective method of hydrogen storage without any risk in the future.
- ii. **Solar Photovoltaic cells:** Solar cells convert the abundantly available sunlight in to electricity without any pollution. However at present the disadvantage of silicon based solar cells is their lower efficiency and higher cost. The use of nanomaterials in

the solar cells will increase the efficiency of the solar cells and reduce their size. Dye sensitized solar cells based on nanoparticles of TiO_2 and dye molecules are cost effective. Various hybrid solar cells based on nanoparticles of CdS , CdSe or ZnO , Carbon Nano Tubes (CNTs), graphene and polymers are under intense research. Physicists expect that in future, such solar cells can have efficiency up to 90%!

- iii. **Batteries:** Several electronic gadgets in the modern life, such as mobile phones, laptops, radio, CD players, toys and watches require rechargeable batteries. The batteries being used at present have low energy density and low storage capacity. This requires frequent charging and frequent replacements. The life of conventional batteries is low. The improved batteries based on nanoparticles of nickel hydrides, and aerogels will overcome the limitations of the present batteries. These materials will be used for electrodes in the batteries.
- c. **Automobiles:** Automobile sector will be greatly benefitted by nanotechnology. There are several areas in the automobile sector which will be receiving these benefits. These are parts of automobile, painting, fuel, motors, window glasses and controlling the pollution.
 - i. **Parts of the automobile:** The structure of the automobile makes the use of several materials such as steel, some alloys, rubbers, plastics etc. The body structure of the car should be strong, rigid (non-deformable) with desirable size and shape. It has been established that the composites based on carbon nanotubes (CNT) are stronger than steel. However, at present CNTs are not commercially feasible. However attempts are being made to reduces the cost. Thus in future, the conventional materials in the body structure will be replaced by composites of CNTs
 - ii. **Painting:** Cars are coated by spray painting. The paints based on the nanoparticles provide smooth, thin and attractive coating. In future it will be possible to tune the colors of the car by applying a small voltage.
 - iii. **Window glasses:** At present window glasses require frequent cleaning due to deposition of the dust. These glasses also become hazy in the rainy season due to water droplets. Both these problems can be overcome by using nanotechnology. Nanotechnology provides self-cleaning photo-catalytic and hydrophilic glasses. 'Self-cleaning glasses' in which glass is doped by the nanoparticles of titania (TiO_2) will replace conventional glasses. Due to TiO_2 , the deposition of the dust and water on the glass is prevented. The dust is decomposed and water droplets are spread uniformly in such glasses.
 - iv. **Motors:** The motors required in the various functions of the car such as wipers, window glass movements, removing CD players etc. The motors based on the nanoparticles of Ni-Ti are more powerful and they consume require less electrical power and thus they will replace the conventional motors.
 - v. **Tyres:** At present the tyres of the motors based on the rubber are bulky and heavy and thus they consume more rubber. This increases the cost, fuel consumption and limits the

speed. The tyres based on nanoclay will be thinner, light weight and less rubber consuming.

vi. Pollution control: The heavy traffic results in to pollution due to emission of poisonous gases like CO, NO etc. The efficient nanocatalysts can convert the harmful emissions in to less harmful gases.

vii. Fuel: Hydrogen can be used as a fuel and can replace petrol and diesel. Hydrogen can be easily obtained from water and thus is abundantly available. Unlike oil and petrol and diesel, it doesn't cause pollution. However, due to its highly combustible nature, it easily catches fire and therefore its storage is difficult. The storage devices based on carbon nanotubes will overcome this problem.

d. Textiles: The fibers, threads and dyes based on nanotechnology are coming up. If these are used in the textiles, then a textile having an attractive look of synthetic fiber but the comfort of the cotton is possible. Such textiles will be self-cleaning and wrinkle free and thus they will require less-frequent cleaning and ironing. The washing machines based on silver nanoparticles are coming up. Due to antibacterial properties of silver, the clothes washed in such machines will be remaining germ free for a longer period.

e. Cosmetics: The nanoparticles of zinc oxide and titanium oxide absorb ultraviolet radiations strongly. Thus the sunscreen lotions based on these nanoparticles can protect the skin from UV radiations. The creams based on these nanoparticles are required in fewer amounts and unlike ordinary creams they do not give whitish tinge. These creams scatter the light in such a way that the skin appears wrinkle free. Further, these creams spread uniformly and leave no gaps. The dyes and colors based on nanoparticles are quite harmless.

f. Medical:

- i. **Cancer diagnosis and therapy:** In chemotherapy, along with cancer cells, the healthy cells are also destroyed. This can be avoided by targeted drug delivery using nanocapsules. The nanocapsules containing the drug can be targeted towards the specific part of the body and then it can be opened at the specific place at a desired rate by using magnetic field or infrared radiation. Nanotechnology also offers early detection of cancer before it is too late.
- ii. In future HIV and diabetes can be treated by using nanomedicines .
- iii. Biocompatible and strong body implants are possible using nanotechnology
- iv. The sensors based on porous silicon and carbon nanotubes offer quick detection of viruses, DNA, protein and antibody.

g. Sports and Toys:

- i. **Tennis balls:** The life of ordinary tennis balls is limited due to leakage of the air through the pores. In the tennis balls made using nanoclay, the pores are filled and hence the leakage of air is avoided. The balls thus have increased life.
- ii. **Rackets:** The use of carbon nanotubes in tennis rackets makes them strong, and lightweight
- iii. **Toys:** Nanotechnology based motors are used in the toys

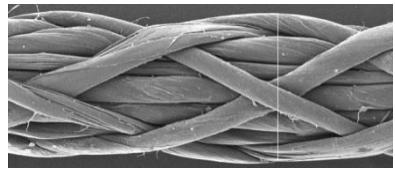
h. Space and Defense:

Due to their superior properties, nanoparticles are replacing conventional materials in space and defense.

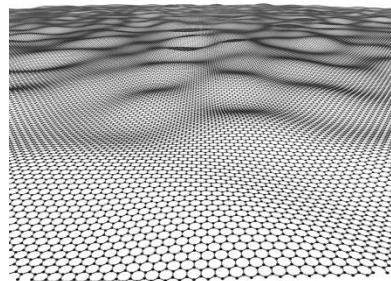
- 1. **Aerogels:** Aerogels are the materials which contain nanopores. Aerogels can be made using different materials. Due to this they are porous and thus have extremely low density as compared to conventional materials. The density of aerogels is almost 20 times smaller than the conventional materials. Thus the weight of aerogels is significantly low as compared to conventional materials. Further, aerogels are poor conductors of heat. This makes them useful in defense and spacecraft where light weight is desirable. Aerogels can also be used in light weight suits and jackets.
- 2. **Solar cells in the spacecrafts:** The solar cells based on nanomaterials have higher efficiency, lower size and lower weight as compared to conventional ones. This is highly beneficial in space crafts, which are invariably powered using solar panels.
- 3. **Special materials for space vehicles:** The space missions have to withstand the extreme environmental conditions, especially during launching and reentry such as exposure to high energy radiations, extreme temperature and pressure etc. Polymer composites doped with silica fibers or nanoparticles have higher Young's modulus, high impact strength and low temperature coefficient of expansion. This makes them suitable for spacecrafts. The polymers doped with nanoparticles can withstand the high energy radiations rather than the polymers based with microparticles. Nanotechnology also promises its role in detecting the biological weapons, decomposition of warfare chemicals, or making the objects invisible.

Carbon based Nanoparticles (an aside)

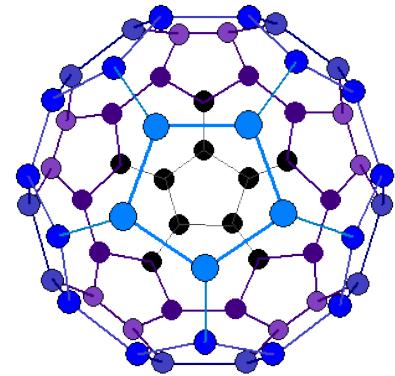
- i. **Carbon nanotubes**
- ii. **Fullerenes**
- iii. **Graphene**
- iv. **Quantum dots**



Carbon Nanotubes



Graphene



Buckminsterfullerene
(C₆₀)

Nanotechnology in the Future

“Richard Feynman’s prophetic speech in 1959 in front of American Physical Society.

There is plenty of room at the bottom

http://www.pa.msu.edu/~yang/RFeynman_plentySpace.pdf



Washing machine (Samsung) based on silver nanoparticles. Silver nanoparticles are used to kill the bacteria

REFERENCE BOOKS

1. Nanotechnology: Principles and Practices by Sulbha Kulkarni, Capital Publishing Co. New Delhi.
2. Introduction to nanotechnology, by C. P. Poole Jr. and F. J. Ownes, Willey Publications.
3. Origin and development of nanotechnology by P. K. Sharma, Vista International publishing house.
4. Nanostructure and nanomaterials synthesis, Properties and applications, by Guozhong Cao, Imperials College Press, London.
5. Nanomaterials: Synthesis, Properties & Applications. Edited by A.S. Edelstein & R.C. Commorata. Institute of Physics Publishing, Bristol & Philadelphia.
6. Nano: The Essentials. T. Pradeep , McGraw Hill Education.
7. Nanotechnology: Fundamentals and applications by Manasi Karkare, I.K. International Pvt. Ltd, New Delhi (2008).

WORLD WIDE WEB

1. “Richard Feynman’s prophetic speech in 1959 in front of American Physical Society.
There is plenty of room at the bottom
http://www.pa.msu.edu/~yang/RFeynman_plentySpace.pdf
2. www.nanotech-now.com/
3. www.understandingnano.com
4. <https://www.nanowerk.com>