GURUNANAK INSTITUTE OF TECHNOLOGY

157/F, Nilgunj Road, Panihati
Kolkata -700114
Website: www.gnit.ac.in
Email: info.gnit@jisgroup.org

Approved by A.I.C.T.E., New Delhi
Affiliated to MAKAUT, West Bengal

OCW

Solid State Devices

Course Level: Undergraduate

Credit: 3

Prepared by:

Dr. Sunipa Roy
Associate Professor (ECE)

Ms. Antara Ghosal
Assistant Professor (ECE)

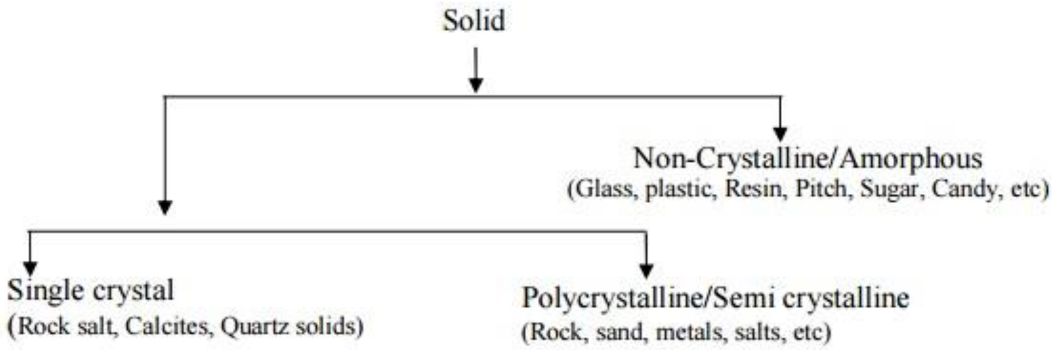
Module I: Energy Band Theory, Charge Carriers in Semiconductors:

Crystalline and Non-crystalline solids:

This can be subdivided into two states-solid and fluid, of which the later is subdivided into liquid and gaseous state. It can also be subdivided into condensed stated and gaseous state where condensed state is subdivided into the solid and liquid state. Although very little of the matter in the universe is in the solid state, solids constitute much of the physical world around us and a large part of the modern technology is based on the special characteristics of the various solid materials.

Crystalline and non-crystalline (Amorphous Solids):

Nature favors the crystalline state of the solids, because the energy of the ordered atomic arrangement is lower than that of an irregular packing of atoms.



Crystalline Solids:

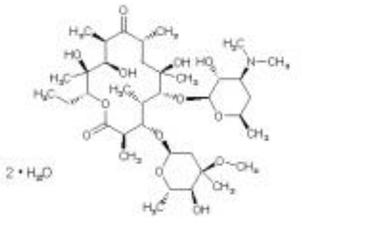
A solid in general is said to be a crystal if the constituent particles (atoms, ions or molecules) are arranged in a three dimensional periodic manner or simply it has a reticular structure. In crystalline solids the atoms are stacked in a regular manner, forming a 3-D pattern which may be obtained by a 3-D repetition of a certain pattern unit. It has long-range orderness and thus has definite properties such a sharp melting point. Thus we can say, crystal is a three dimensional periodic array of atoms. When the crystal grows under constant environment, the external geometrical shape of the crystal often remains unchanged. Thus, the shape is a consequence of the internal arrangement of constituent particles. The ideal crystal has an infinite 3D repetition of identical units, which may be atoms or molecules. All ionic solids and most covalent solids are crystalline. All solid metals, under normal circumstances, are crystalline.

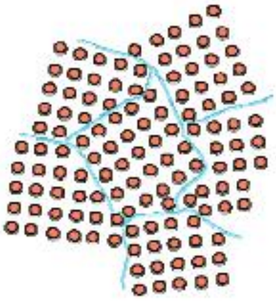
Single crystal :

Solid Non-Crystalline/Amorphous (Glass, plastic, Resin, Pitch, Sugar, Candy, etc) Single crystal (Rock salt, Calcites, Quartz solids) Polycrystalline/Semi crystalline (Rock, sand, metals, salts, etc) Crystalline and Non-crystalline solids 2 When the periodicity in crystal pattern extends throughout a certain piece of materials, one speaks of a single crystal or unit crystal or mono-crystal. Rock salt, calcites, quartz, etc. are examples of common single crystal.

Polycrystalline solids (Polymorphism):

When the periodicity in the crystal structure is interrupted at so-called grain boundaries, the crystal is said to be polycrystalline. In this case the size of the grains or crytallites is smaller than the size of the pattern unit which forms the periodicity. The size of the grain in which the structure is periodic may vary from macroscopic dimensions to several angstroms. In general, the grains in such a solid are not related in shape to the crystal structure, the surface being random in shape rather than well defined crystal planes. Rock, sand, metals, salts, etc. are some examples of polycrystalline solids.





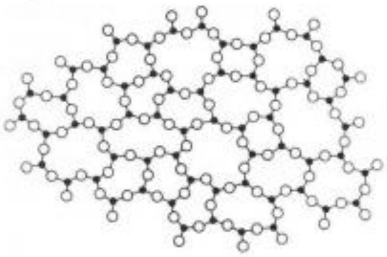


Figure 1: Single crystal, polycrystalline, amorphous[1]

Noncrystalline solids :

It is the opposite extreme of a single crystal. These types of solids have neither reticular nor granular structure. At most causes exhibit short range orderness in their structure. Glass and plastic are common example of this class. When the size of the grains or crystallites becomes comparable to the size of the pattern unit, we speak of amorphous substances. A typical feature of these substances is that they have no definite melting points. As their temperature is increased, they gradually become soft; their viscosity drops, and begins to behave like ordinary viscous liquids. Amorphous solids have no long-range order. The atoms or molecules in these solids are not periodically located over large distances. An amorphous structure is shown below.

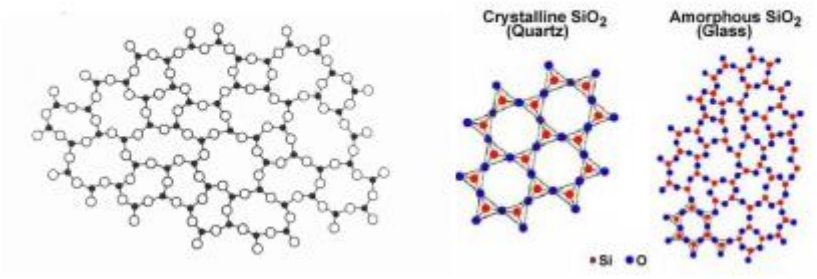


Figure 2: Different crystal structures[1]

Many amorphous materials have internal structures similar to liquids. In fact, the only obvious distinction between amorphous materials, such as glass, and liquids is the high viscosity (resistance to flow) of the amorphous solids. All solids tend to exist in the crystalline state rather than the amorphous state because the crystalline structure always has a larger binding energy. However, in numerous instances amorphous solids are formed when liquids are cooled below the melting temperature.

This occurs for two reasons:

1) The structure of the molecules is so complex that they cannot easily rearrange themselves to form a crystalline structure, and/or

2) The solid forms so rapidly that the atoms or molecules do not have time enough to rearrange themselves in a crystalline structure.

Generally, amorphous solids have one of two distinct atomic arrangements: either a tangled mass of long-chained molecules or a 3-dimentional network of atoms with no long-range order. Amorphous materials with long-chained molecules (e.g. polymers) have a structure similar to that shown below.

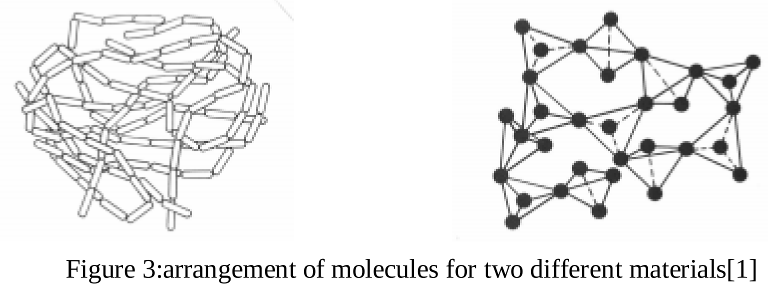


Figure 3:arrangement of molecules for two different materials[1]

Each segment in above figure represents one of the repeating units of the polymer chain. The arrangement of the molecules is fairly random, resulting in a loosely packed structure. Network amorphous solids are usually Oxides, the most common being Silica (SiO2). The amorphous SiO2 structure is also shown above. Only oxygen atoms are shown (corners of tetrahedral) in this amorphous SiO2 structure. There is a Silicon atom at the center of each tetrahedral which is not shown here. This structure has short-range order but none of the long-range order found in crystalline Silica. Thus, in both amorphous and crystalline Silica, each Silicon atom and each Oxygen atom have essentially the same local surroundings, even though there is no long-range periodicity in the amorphous structure.

Solids that do not have long range atomic order are called amorphous solids. They often have subunits that have consistent form, but their long-range order is disturbed because the sub-units pack randomly. Amorphous solids are formed when liquids are cooled too quickly from the molten state to allow the sub-units to arrange themselves in the low energy, crystalline state. Solids with pure ionic bonds do not form amorphous solids but all the other bond types can produce amorphous solids. Silica (SiO2) can form either covalent amorphous solids, usually called glasses or regular crystal structures (Quartz).

Crystal Planes and Miller Indices:

Index system for crystal directions and planes Crystal directions: The direction is specified by the three integers [n1n2n3]. If the numbers n1n2n3 have a common factor, this factor is removed. For example, [111] is used rather than [222], or [100], rather than [400]. When we speak about directions, we mean a whole set of parallel lines, which are equivalent due to transnational symmetry. Opposite orientation is denoted by the negative sign over a number. For example:

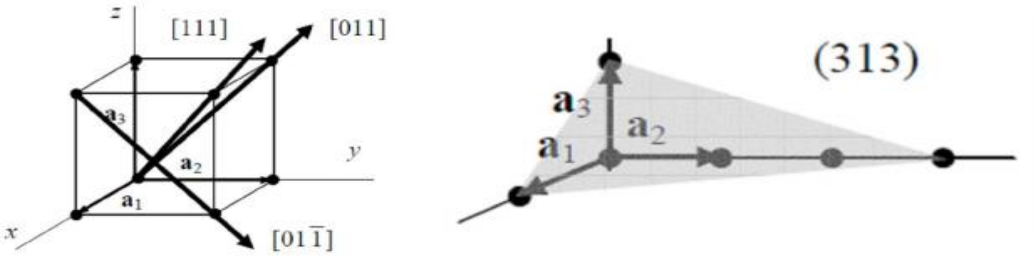


Figure 4: Crystal planes

Crystal planes:

The orientation of a plane in a lattice is specified by Miller indices. They are defined as follows. We find intercept of the plane with the axes along the primitive translation vectors a1, a2 and a3. Let's these intercepts be x, y, and z, so that x is fractional multiple of a1, y is a fractional multiple of a2 and z is a fractional multiple of a3. Therefore we can measure x, y, and z in units a1, a2 and a3 respectively. We have then a triplet of integers (x y z). Then we invert it (1/x 1/y 1/z) and reduce this set to a similar one having the smallest integers by multiplying by a common factor. This set is called Miller indices of the plane (hkl). For example, if the plane intercepts x, y, and z in points 1, 3, and 1, the index of this plane will be (313). The orientation of a crystal plane is determined by three points in the plane, provided they are not collinear. If each point lay on a different crystal axis, the plane could be specified by giving the coordinates of the points in terms of the lattice constants a, b, c. A notation conventionally used to describe lattice points (sites), directions and planes is known as Miller Indices. A crystal lattice may be considered as an assembly of equidistant parallel planes passing through the lattice points and are called lattice planes. In order to specify the orientation one employs the so called Miller indices. For simplicity, let us start with a two dimensional lattice and then generalized to three dimensional case. The equation of plane in 2-D and 3D having the intercepts a, b and a, b, c respectively are

$(x/a) + (y/b) =1$

And

$(x/a) + (y/b) +(z/c) = 1$

Crystal direction is the direction (line) of axes or line from the origin and denoted as [111], [100], [010] etc.

Schrodinger Wave Equation :

The free electron model gives us a good insight into many properties of metals, such as the heat capacity, thermal conductivity and electrical conductivity. However, this model fails to help us other important properties. For example, it does not predict the difference between metals, semiconductors and insulators. It does not explain the occurrence of positive values of the Hall coefficient. Also the relation between conduction electrons in the metal and the number of valence electrons in free atoms is not always correct. We need a more accurate theory, which would be able to answer these questions. The problem of electrons in a solid is in general a many-electron problem. The full Hamiltoniam of the solid contains not only the one-electron potentials describing the interactions of the electrons with atomic nuclei, but also pair potentials describing the electron-electron interactions. The many-electron problem is impossible to solve exactly and therefore we need simplified assumptions. The simplest approach we have already considered, it is a free electron model. The next step in building the complexity is to consider an independent electron approximation, assuming that all the interactions are described by an effective potential. One of the most important properties of this potential is that it is periodic on a lattice

$$U(r) =U(r + T) \qquad (1)$$

where T is a lattice vector. Qualitatively, a typical crystalline potential might be expected to have a form shown in Fig.1, resembling the individual atomic potentials as the ion is approached closely and flattening off in the region between ions.

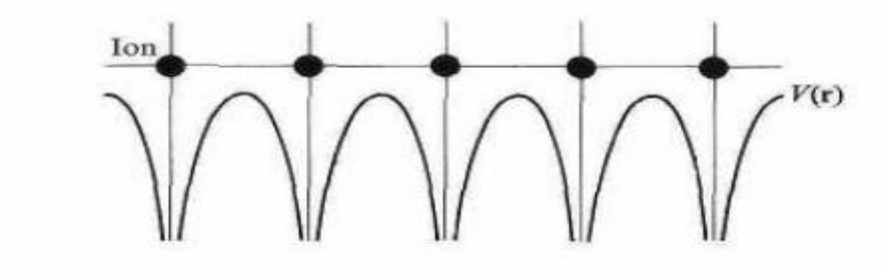


Figure 5: The crystal potential seen by the electron

Within the approximation of non-interacting electrons the electronic properties of a solid can be examined by Schrödinger equation

$$\left[-\frac{\hbar^2}{2m}\nabla^2+U(\mathbf{r})\right]\psi(\mathbf{r})=E\psi(\mathbf{r})$$

(2)

in which ψ (r) is a wave function for one electron. Independent electrons, which obey a one electron Schrödinger equation (2) with a periodic potential, are known as Bloch electrons, in contrast to "free electrons," to which Bloch electrons reduce when the periodic potential is identically zero.
Now we discuss general properties of the solution of the Schrödinger equation (2) taking into account periodicity of the effective potential (1) and discuss main properties of Bloch electrons, which follow from this solution. We represent the solution as an expansion over plain waves:

$$\psi(\mathbf{r})=\sum_{\mathbf{k}}c_{\mathbf{k}}e^{i\mathbf{kr}}$$

(3)

This expansion in a Fourier series is a natural generalization of the free-electron solution for a zero potential. The summation in (3) is performed over all k vectors, which are permitted by the periodic boundary conditions. According to these conditions the wave function (3) should satisfy

ψ (x, y,z) =ψ (x + L, y,z) =ψ (x, y + L,z) =ψ (x, y,z + L) (4)

so that

$$k_x=\frac{2\pi n_x}{L};\quad k_y=\frac{2\pi n_y}{L};\quad k_z=\frac{2\pi n_z}{L}$$

(5)

where nx, ny, and nz are positive or negative integers. Note that in general ψ (r) is not periodic in the lattice translation vectors. On the other hand, according to Eq.(1) the potential energy is periodic, i.e. it is invariant under a crystal lattice translation. Therefore, its plane wave expansion will only contain plane waves with the periodicity of the lattice. Therefore, only reciprocal lattice vectors are left in the Fourier expansion for the potential:

$$U(\mathbf{r})=\sum_{\mathbf{G}}U_{\mathbf{G}}e^{i\mathbf{Gr}}$$

(6)

where the Fourier coefficients UG are related to U(r) by

$$U_{\mathbf{G}}=\frac{1}{V_c}\int_{cell}e^{-i\mathbf{Gr}}U(\mathbf{r})dr$$

(7)

where Vc is the volume of the unit cell. It is easy to see that indeed the potential energy represented by (6) is periodic in the lattice:

$$U(\mathbf{r}+\mathbf{T})=\sum_{\mathbf{G}}U_{\mathbf{G}}e^{i\mathbf{G}(\mathbf{r}+\mathbf{T})}=e^{i\mathbf{GT}}\sum_{\mathbf{G}}U_{\mathbf{G}}e^{i\mathbf{Gr}}=U(\mathbf{r})$$

(8)

where the last equation comes from the definition of the reciprocal lattice vectors eiGT = 1 . The values of Fourier components UG for actual crystal potentials tend to decrease rapidly with increasing magnitude of G. For example, for a Coulomb potential UG decreases as 1/G2 . Note that since the potential energy is real the Fourier components should satisfy U−G =U*G .

We now substitute (3) and (6) in Eq.(2) and obtain:

$$\frac{\hbar^2}{2m}\sum_{\mathbf{k}}k^2c_{\mathbf{k}}e^{i\mathbf{kr}}+\sum_{\mathbf{k}}\sum_{\mathbf{G}}U_{\mathbf{G}}c_{\mathbf{k}}e^{i(\mathbf{k}+\mathbf{G})\mathbf{r}}=E\sum_{\mathbf{k}}c_{\mathbf{k}}e^{i\mathbf{kr}}$$

(9)

Changing the summation index in the second sum on the left from k to k+G this equation can be rewritten in a form:

$$\sum_{\mathbf{k}}e^{i\mathbf{kr}}\left\{\left(\frac{\hbar^2}{2m}k^2-E\right)c_{\mathbf{k}}+\sum_{\mathbf{G}}U_{\mathbf{G}}c_{\mathbf{k}-\mathbf{G}}\right\}=0$$

(10)

Since this equation must be satisfied for any r the Fourier coefficients in each separate term of (10) must vanish and therefore

$$\left(\frac{\hbar^2}{2m}k^2-E\right)c_{\mathbf{k}}+\sum_{\mathbf{G}}U_{\mathbf{G}}c_{\mathbf{k}-\mathbf{G}}=0$$

(11)

This is a set of linear equations for the coefficients ck. These equations are nothing but restatement of the original Schrödinger equation in the momentum space, simplified by the fact that the potential is periodic. This set of equations does not look very pleasant because, in principle, an infinite number of coefficients should be determined. However, a careful examination of Eq.(11) leads to important consequences.

First, we see that for a fixed value of k the set of equations (11) couples only those coefficients, whose wave vectors differ from k by a reciprocal lattice vector. In the one-dimensional case these are k, k±2π/a, k±4π/a, and so on. We can therefore assume that the k vector belongs to the first Brillouin zone. The original problem is decoupled to N independent problems (N is the total number of atoms in a lattice): for each allowed value of k in the first Brillouin zone. Each such problem has solutions that are superposition of plane waves containing only the wave vector k and wave vectors differing from k by the reciprocal lattice vector. Putting this information back into the expansion (3) of the wave function ψ (r) , we see that the wave function will be of the form

$$\psi_{\mathbf{k}}(\mathbf{r})=\sum_{\mathbf{G}}c_{\mathbf{k}-\mathbf{G}}e^{i(\mathbf{k}-\mathbf{G})\mathbf{r}}$$

(12)

where the summation is performed over the reciprocal lattice vectors and we introduced index k for the wave function. We can rearrange this so that

$$\psi_{\mathbf{k}}(\mathbf{r})=e^{i\mathbf{kr}}\sum_{\mathbf{G}}c_{\mathbf{k}-\mathbf{G}}e^{-i\mathbf{Gr}}$$

(13)

Or

$$\psi_{\mathbf{k}}(\mathbf{r})=e^{i\mathbf{kr}}u_{\mathbf{k}}(\mathbf{r})$$

(14)

where uk(r)= uk(r+T) is a periodic function which is defined by

$$u_{\mathbf{k}}(\mathbf{r}) = \sum_{\mathbf{G}} c_{\mathbf{k}-\mathbf{G}} e^{-i\mathbf{G}\mathbf{r}} \tag{15}$$

Equation (14) is known as Bloch theorem, which plays an important role in electronic band structure theory.

Kronig–Penney Model :

The fundamental nature of insulators, conductors and semiconductors can be functionally explained based on band theory. The recent development in semiconductor physics, the semiconductor hetero-structures are also analyzed using the concept of band theory. Another notable theory, the free-electron theory, can help in understanding the electron movement in metals. It assumes that, the valence electron in a metal absorbs thermal energy which ultimately is converted into kinetic energy with an average of (3/2)KBT based on law of equipartition of energy. But the calculated molar electronic specific heat does not match with the experimental value. Hence it can be concluded that the equipartition law and the classical MaxwellBoltzmann statistics are not adequate for evaluating electronic specific heat in metals. Another failure of classical free electron theory is that, it does not account for the magnetic moment of electron due to its spin.

The free-electron theory, which neglects the magnetic moment of electrons arising from their spin predicts that, paramagnetic susceptibility is proportional to the temperature for each electron. On the contrary, the experimental results show that the susceptibility is almost independent (constant) of temperature. The reason is, the classical theory allows all the free electrons to gain energy which does not actually happen in reality, which leads to drastic difference between the calculated and the observe values. At this juncture, the quantum free-electron theory steps in, assumes that an electron in a metal experiences a constant or zero potential and hence is free to move within the lattice. The quantum free-electron theory thus successfully explains the specific heat, electrical conductivity, thermionic emission, thermal conductivity and para magnetism of materials. However, the concept fails to differentiate the conductivities in conductors, semiconductors and insulators. In a real crystal, electrons move in a regularly arranged lattice of positive ions. The electrons have the zero potential at the positive ion site and possess maximum value at the intermediate lattice points. This could be schematically represented as shown in Fig. 2(sine wave notation). The observed potential is periodical as the lattice planes. Bloch has the solution as ψ (x) = Uk (x)e ikx for the Schrödinger equation, which describes the electron motion:

$$\frac{d^2\psi}{dx^2} + \frac{2m}{\hbar^2}[E - V(x)]\psi = 0$$

where, ψ is the wave function and Uk(x) in the Bloch's solution, which has the periodicity of the lattice.

Hence, the wave function includes both a plane wave eikx which is modulated by the periodic function Uk(x) and the state of motion of electron, which is represented by the wave vector k. However, it is difficult (not tractable) to solve the Schrödinger's equation with the sinusoidal periodicity. Therefore, Kronig and Penney suggested a simpler model, where the inner potential of the crystal system has the rectangular shaped potential.

Thus, in the Kronig Penney model, instead of experiencing a gradual variation in the strength of the potential electrons experience a maximum potential (potential well) and minimum value (potential barrier) in the presence of the lattice planes.

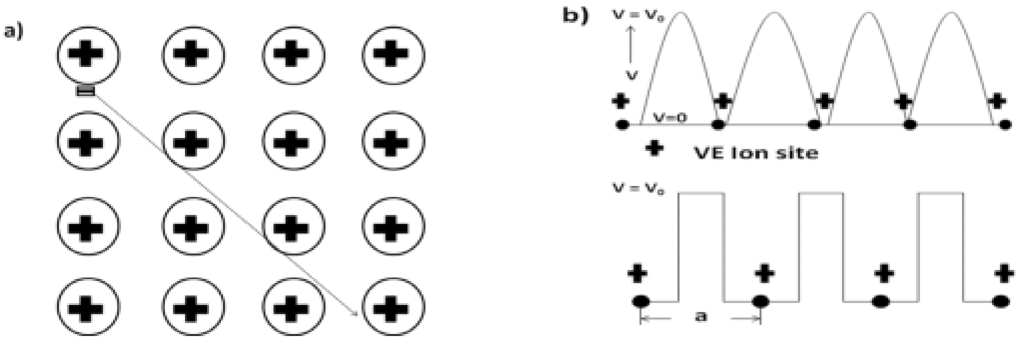
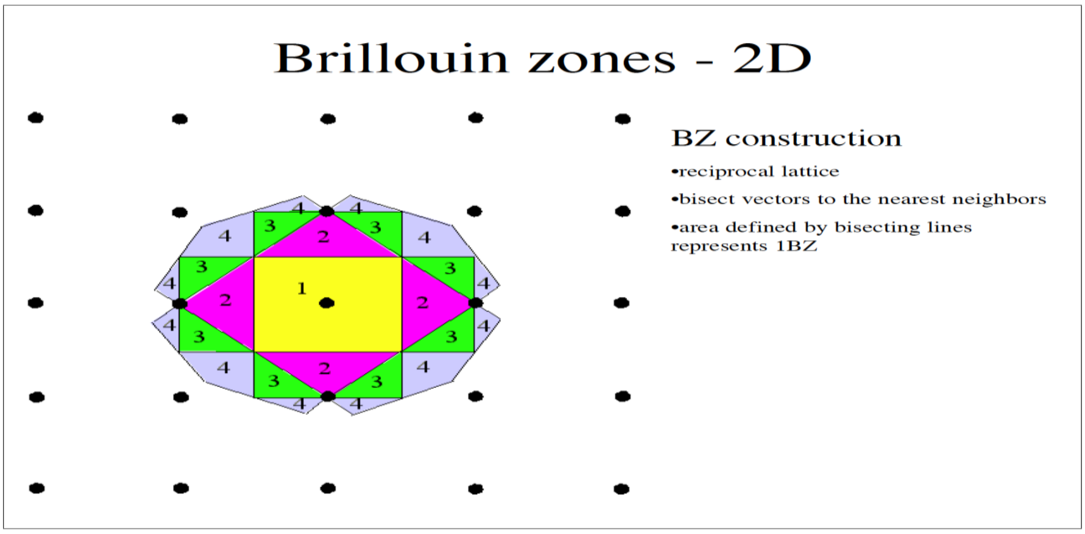


Figure 6: One dimensional periodic potential distribution for a crystal

Two Dimensional Brillouin Zone:



Brillouin zones - 2D

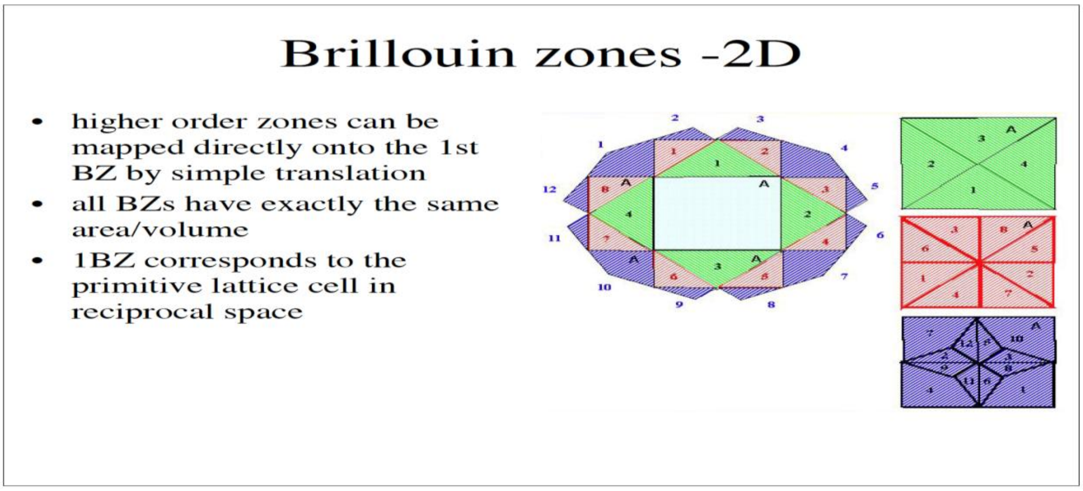
BZ construction

- reciprocal lattice
- bisect vectors to the nearest neighbors
- area defined by bisecting lines represents 1BZ

Figure 7: Brillouin zone

We construct the first Brillouin zone from the shortest lattice vector G_1as follows.
We construct the second Brillouin zone from the next shortest vector G_2 and so on.

Brillouin Zones- 2D



Brillouin zones -2D

- higher order zones can be mapped directly onto the 1st BZ by simple translation
- all BZs have exactly the same area/volume
- 1BZ corresponds to the primitive lattice cell in reciprocal space

Figure 8: Brillouin zone

construction

- Reciprocal lattice
- Bisect vectors to the nearest neighbors
- Area defined by bisecting lines represents 1BZ

Three Dimensional Brillouin Zones:

- A 3-dimensional Brillouin zone can be constructed in a similar way by bisecting all lattice vectors and placing planes perpendicular to these points of bisection.
- This is similar to the Wigner Seitz cell in the real lattice.

Wigner Seitz Cell:

- A primitive unit cell which shows the cubic symmetry of the lattice(for the cubic system).(real lattice)
- The First Brillouin zone is the Wigner Seitz cell in the reciprocal lattice

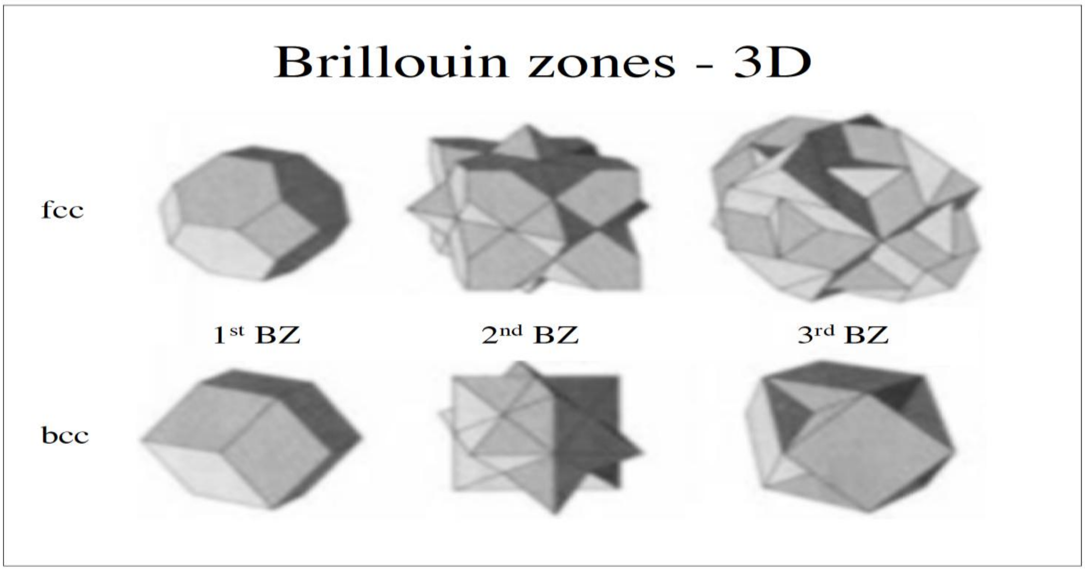


Figure 9 : Brillouin zone

Lets Study These Figures:

- *First Brillouin zone of the bcc structure
- ⇒Free electron bands for bcc structure
- *First Brillouin zone of the fcc structure
- ⇒Free electron bands for fcc structure
-

Explanation of these symbols:
Look between the graph of bands and the first Brillouin zone, you will find:

$$\Gamma\text{: center of the Brillouin zone}$$
$$\text{X: [100]intercept}$$
$$\text{K: [110]intercept}$$
$$\text{L: [111]intercept}$$
$$\Gamma - \text{X: path } \Delta$$
$$\Gamma - \text{L: path } \Lambda$$
$$\Gamma - \text{K: path } \Sigma$$

Number of states in the band:

In solid-state and condensed matter physics, the density of states (DOS) of a system describes the number of states per interval of energy at each energy level that are available to be occupied.

Band gap in the nearly free electron model:

Having derived Bloch's theorem we are now at a stage where we can start introducing the concept of bandstructure. When someone refers to the bandstructure of a crystal they are generally talking about its electronic dispersion, E(k) (i.e. how the energy of an electron varies as a function of crystal wavevector). However, Bloch's theorem is very general and can be applied to any periodic interaction, not just to electrons in the periodic electric potential of ions. For example in recent years the power of band theory has been applied to photons in periodic dielectric media to study photonic bandstructure (i.e. dispersion relations for photons in a "photonic crystal"). In this lecture we will firstly take a look at dispersion for an electron in a periodic potential where the potential very weak (the nearly free electron approximation) and in the next lecture we will look at the case where the potential is very strong (tight binding approximation). Firstly let's take a closer look at dispersion.

You will recall from the Sommerfeld model that the dispersion of a free electron is E(k) = (h¯ 2k 2)/(2m) . It is completely isotropic (hence the dispersion only depends on k = |k|) and the Sommerfeld model produces exactly this band structure for every material – not very exciting! Now we want to understand how this parabolic relation changes when you consider the periodicity of the lattice. Using Bloch's theorem you can show that translational symmetry in real space (characterised by the set translation vectors {T}) leads to translational symmetry in k-space (characterised by the set of reciprocal lattice vectors {G}). Knowing this we can take another look at Schr¨odinger's equation for a free electron in a periodic potential V (r) :

$$H\psi_{\nu\mathbf{k}}(\mathbf{r}) = \{-\frac{\hbar^2\nabla^2}{2m} + V(\mathbf{r})\}\psi_{\nu\mathbf{k}}(\mathbf{r}) = E_{\nu\mathbf{k}}\psi_{\nu\mathbf{k}}(\mathbf{r})$$

and taking the limit V (r) → 0 we know that we have a plane wave solution. This implies that the Bloch function u(r) → 1. However considering the translational invariance in k-space the dispersion relation must satisfy:

$$E_{\nu \mathbf{k}} = \frac{\hbar^2|\mathbf{k}|^2}{2m} = \frac{\hbar^2|\mathbf{k}+\mathbf{G}|^2}{2m}$$

for the set of all reciprocal lattice vectors {G}. This dispersion relation is show in Figure 10.

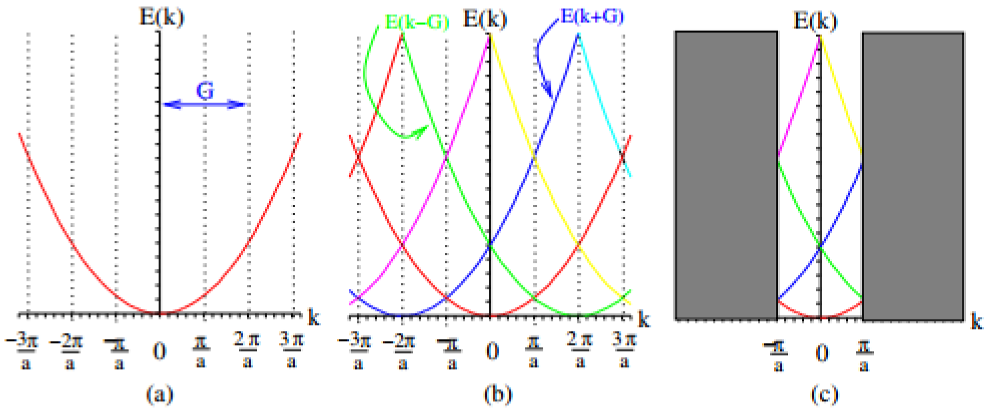


Figure 10: Simple bandstructure diagrams for a one dimensional periodic solid in the limit V (r) → 0 expressed in the extended zone (a), repeated zone (b), and reduced zone (c) schemes.

Since we are in the weak potential limit we can treat the crystal potential as a weak perturbation added to the Hamiltonian of a free electron. Let’s start with the Schr¨odinger equation for a free electron

Hˆ oψvk(r) = Evkψvk(r)

Where

Hˆ o = ˆp 2/(2m)

which has plane wave eigenstates

$$\psi_{i\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{V_{r3}}}\exp(i\mathbf{k}\cdot\mathbf{r})$$

We now introduce a small perturbation, Hˆ 0 associated with the crystal potential

$$\hat{\mathrm{H}} = \hat{\mathrm{H}}_o + \hat{\mathrm{H}}'$$

Since the lattice is periodic we may expand the perturbation into a Fourier series where {G} are a set of vectors and VG are Fourier coefficients

$$\hat{\mathrm{H}}' = V(\mathbf{r}) = \sum_{\{\mathbf{G}\}} V_{\mathbf{G}} \exp(-i\mathbf{G}\cdot\mathbf{r})$$

Since the lattice is periodic we may expand the perturbation into a Fourier series where G are a set of vectors and VG are Fourier coefficients.

Tight binding model:

Solution of the tight binding model is periodic in k. Apparently have an infinite number of k states for each allowed energy state.

In fact the different k states all equivalent Bloch States

$$\psi(\mathbf{r}+\mathbf{R}) \equiv e^{i\mathbf{k}.\mathbf{R}}\psi(\mathbf{r})$$

Let k = ḱ ' + G where k′ is in the first Brillouin zone an d G is a rec iproca l latt ice vector.

$$\psi(\mathbf{r}+\mathbf{R}) \equiv \mathrm{e}^{i\mathbf{k}'.\mathbf{R}}\mathrm{e}^{i\mathbf{G}.\mathbf{R}}\psi(\mathbf{r})$$

But G.R = 2 π n, n-integer. Definition of the reciprocal lattice. So

$$\mathrm{e}^{i\mathbf{G}.\mathbf{R}} = 1 \quad \text{and } \psi(\mathbf{r}+\mathbf{R}) \equiv \mathrm{e}^{i\mathbf{k}'.\mathbf{R}}\psi(\mathbf{r}) \qquad e^{i\boldsymbol{k}.\boldsymbol{R}} \equiv e^{i\boldsymbol{k}'.\boldsymbol{R}}$$

k ′ is exactly equivalent to k.

The only independent values of k are those in the first Brillouin zone.

Formation of allowed and forbidden energy bands:

A crystal is a solid consisting of a regular and repetitive arrangement of atoms or molecules (strictly speaking, ions) in space. If the positions of the atoms in the crystal are represented by points, called lattice points, we get a crystal lattice. The distance between the atoms in a crystal is fixed and is termed the lattice constant of the crystal. To discuss the behaviour of electrons in a crystal, we consider an isolated atom of the crystal. If Z is the atomic number, the atomic nucleus has a positive charge Ze. At a distance r from the nucleus, the electrostatic potential due to the nuclear charge is (in SI units)

$$V(r) = \frac{Ze}{4\pi\varepsilon_0 r}$$

where ε0 is the permittivity of free space. Since an electron carries a negative charge, the potential energy of an electron at a distance r from the nucleus is

$$E_p(r) = -eV(r) = -\frac{Ze^2}{4\pi\varepsilon_0 r}$$

V(r) is positive while E p(r) is negative. Both V(r) and E p(r) are zero at an infinite distance from the nucleus. Figs. 1.2(a) and (b) show the variation of V(r) and E p(r), respectively with r.

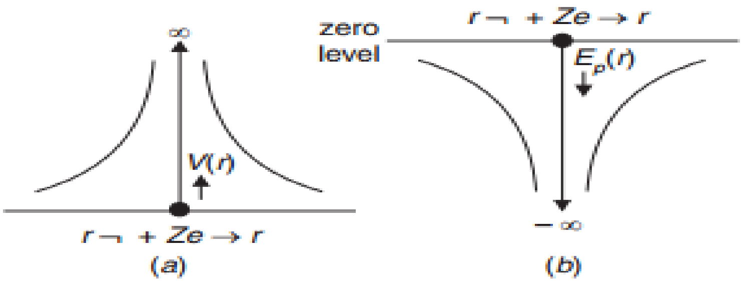


Figure 11: Variation of (a) Potential in the field of a nucleus with distance, (b) Potential energy of an electron with its distance from the nucleus.

We now consider two identical atoms placed close together. The net potential energy of an electron is obtained as the sum of the potential energies due to the two individual nuclei. In the region between the two nuclei, the net potential energy is clearly smaller than the potential energy for an isolated nucleus shown in the figure 11.

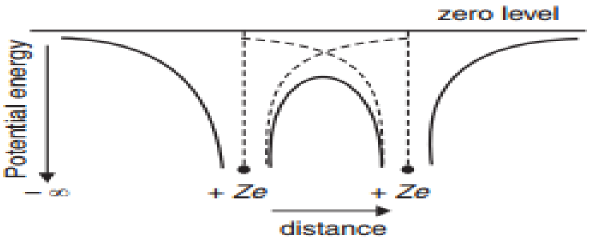


Figure 12: Potential energy variation of an electron with distance between two identical nuclei.

The potential energy along a line through a row of equispaced atomic nuclei, as in a crystal, is diagrammatically shown in Figure 12. The potential energy between the nuclei is found to consist of a series of humps. At the boundary AB of the solid, the potential energy increases and approaches zero at infinity, there being no atoms on the other side of the boundary to bring the curve down.

The total energy of an electron in an atom, kinetic plus potential, is negative and has discrete values. These discrete energy levels in an isolated atom are shown by horizontal lines in Figure. When a number of atoms are brought close together to form a crystal, each atom will exert an electric force on its neighbours. As a result of this interatomic coupling, the crystal forms a single electronic system obeying Pauli's exclusion principle. Therefore, each energy level of the isolated atom splits into as many energy levels as there are atoms in the crystal, so that Pauli's exclusion principle is satisfied. The separation between the split-off energy levels is very small. This large number of discrete and closely spaced energy levels form an energy band. Energy bands are represented schematically by shaded regions in Figure 13.

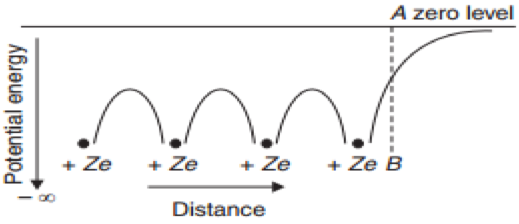


Figure 13: Potential energy of an electron along a row of atoms in a crystal.

The width of an energy band is determined by the parent energy level of the isolated atom and the atomic spacing in the crystal. The lower energy levels are not greatly affected by the interaction among the neighbouring atoms, and hence form narrow bands. The higher energy levels are greatly affected by the interatomic interactions and produce wide bands. The interatomic spacing, although fixed for a given crystal, is different for different crystals. The width of an energy band thus depends on the type of the crystal, and is larger for a crystal with a small interatomic spacing. The width of a band is independent of the number of atoms in the crystal, but the number of energy levels in a band is equal to the number of atoms in the solid. Consequently, as the number of atoms in the crystal increases, the separation between the energy levels in a band decreases. As the crystal contains a large number of atoms (≈ 1029 m–3), the spacing between the discrete levels in a band is so small that the band can be treated as continuous.

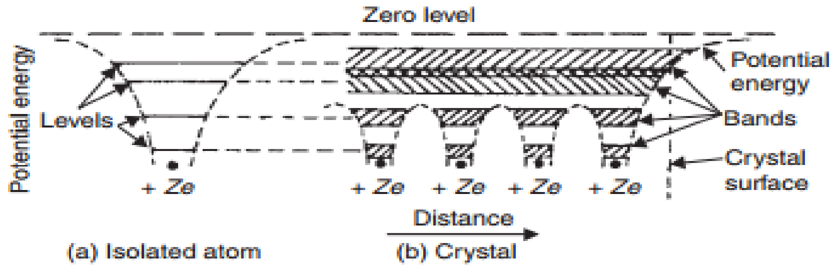


Figure 14: Splitting of energy levels of isolated atoms into energy bands as these atoms are brought close together to produce a crystal.

The lower energy bands are normally completely filled by the electrons since the electrons always tend to occupy the lowest available energy states. The higher energy bands may be completely empty or may be partly filled by the electrons. Pauli's exclusion principle restricts the number of electrons that a band can accommodate. A partly filled band appears when a partly filled energy level produces an energy band or when a totally filled band and a totally empty band overlap.

As the allowed energy levels of a single atom expand into energy bands in a crystal, the electrons in a crystal cannot have energies in the region between two successive bands. In other words, the energy bands are separated by gaps of forbidden energy.

The average energy of the electrons in the highest occupied band is usually much less than the zero level marked in Fig. 1.5(b). The rise of the potential energy near the surface of the crystal, as shown in Fig. 1.5(b), serves as a barrier preventing the electrons from escaping from the crystal. If sufficient energy is imparted to the electrons by external means, they can overcome the surface potential energy barrier, and come out of the crystal surface.

On the basis of the band structure, crystals can be classified into metals, insulators, and semiconductors.

Metal :

A crystalline solid is called a metal if the uppermost energy band is partly filled or the uppermost filled band and the next unoccupied band overlap in energy. Here, the electrons in the uppermost band find neighbouring vacant states to move in, and thus behave as free particles. In the presence of an applied electric field, these electrons gain energy from the field and produce an electric current, so that a metal is a good conductor of electricity. The partly filled band is called the conduction band. The electrons in the conduction band are known as free electrons or conduction electrons.

Insulator:

In some crystalline solids, the forbidden energy gap between the uppermost filled band, called the valence band, and the lowermost empty band, called the conduction band, is very large. In such solids, at ordinary temperatures only a few electrons can acquire enough thermal energy to move from the valence band into the conduction band. Such solids are known as insulators. Since only a few free electrons are available in the conduction band, an insulator is a bad conductor of electricity. Diamond having a forbidden gap of 6 eV is a good example of an insulator. The energy band structure of an insulator is schematically shown in Figure 15.

Semiconductor :

A material for which the width of the forbidden energy gap between the valence and the conduction band is relatively small (~ 1 eV) is referred to as a semiconductor. Germanium and silicon having forbidden gaps of 0.78 and 1.2 eV, respectively, at 0 K are typical semiconductors. As the forbidden gap is not very wide, some of the valence electrons acquire enough thermal energy to go into the conduction band. These electrons then become free and can move about under the action of an applied electric field. The absence of an electron in the valence band is referred to as a hole. The holes also serve as carriers of electricity. The electrical conductivity of a semiconductor is less than that of a metal but greater than that of an insulator. The band diagram of a semiconductor is given in Figure 15.

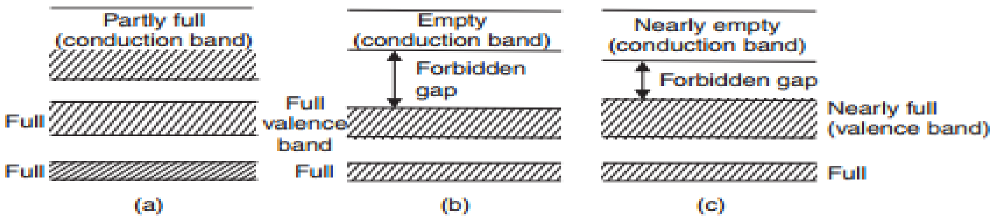


Figure 15: Energy band structure of (a) metal, (b) insulator, and (c) semiconductor. [2]

Effective mass, Wave vector, Energy-band (E-k) diagram: Effective mass (m^*) is the mass that it seems to have when responding to forces or with other identical particles in a thermal distribution. The "wave-particle" motion of electrons in a lattice is not the same as that for a free electron, because of the interaction with the periodic potential of the lattice.

The effective mass is an inverse function of the curvature of the E-k diagram: weak curvature gives large mass, and strong curvature gives small mass.

In general, the effective mass is a tensor quantity, however, for parabolic bands, it is a constant.

Wave vector (k in cm^{-1}) is a variable which states the change of the momentum of an electron in different directions or co-ordinates, while the electron traverse from valance band to conduction band and vice versa.

In a typical quantitative calculation of band structures, the wave function of a single electron traveling through a perfectly periodic lattice is assumed to be in the form of a plane wave moving in the x-direction (say) with propagation constant k, also called a *wave vector*.

The wave vector of the electron is much larger than that of the photon.

It is to be expressed thematically in scalar form, p=(kh)/2*pi, Where p is momentum of electron, h is plunk constant.The momentum changes instantaneously due to the variation of momentum of the electron while travelling through one non-uniform potential barrier.

k = (2*pi)/wavelength this is the expression illustrates the variation of wave vector, due to the generated wave length by the electron.

The band structure of a crystalline solid, i.e allowed values of energy, while plotted as a function of k, gives energy-momentum ***(E-k)*** relationship is usually obtained by solving the Schrodinger equation of an approximate one-electron problem. It shows the relationship between the energy and momentum of available quantum mechanical states for electrons in the material.

Relation between E-K diagram & Effective mass:

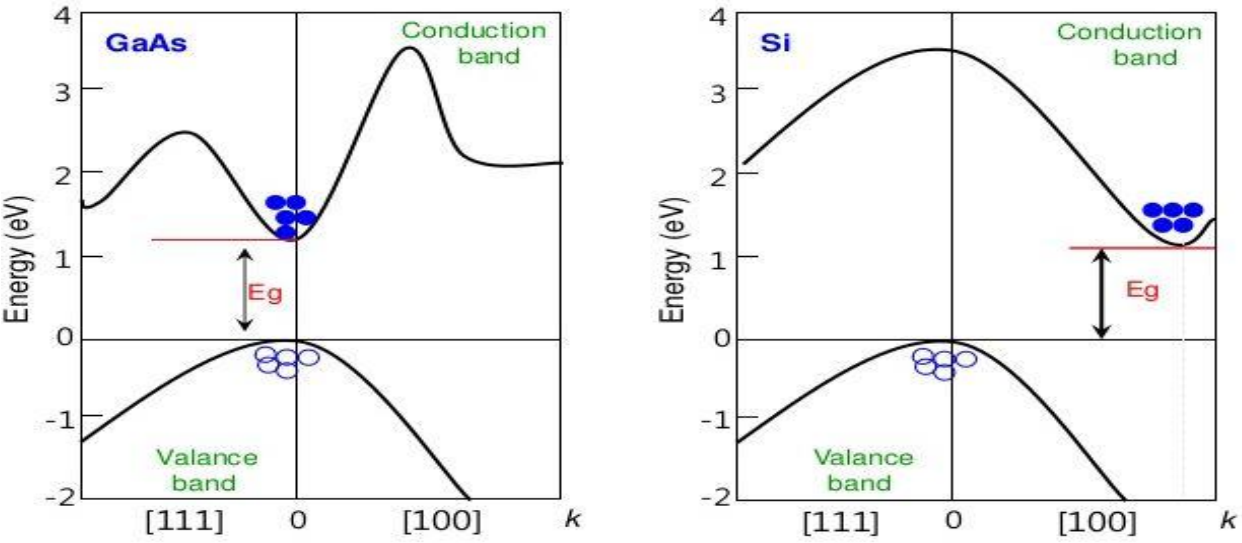


Figure 16. Energy-band (E-k) diagram of GaAs and Si

Debye length: Debye length is the distance over which significant charge separation can occur, is the measure of a mobile charge carrier's (e.g. electrons) screen out electric fields in plasmas and other conductors.

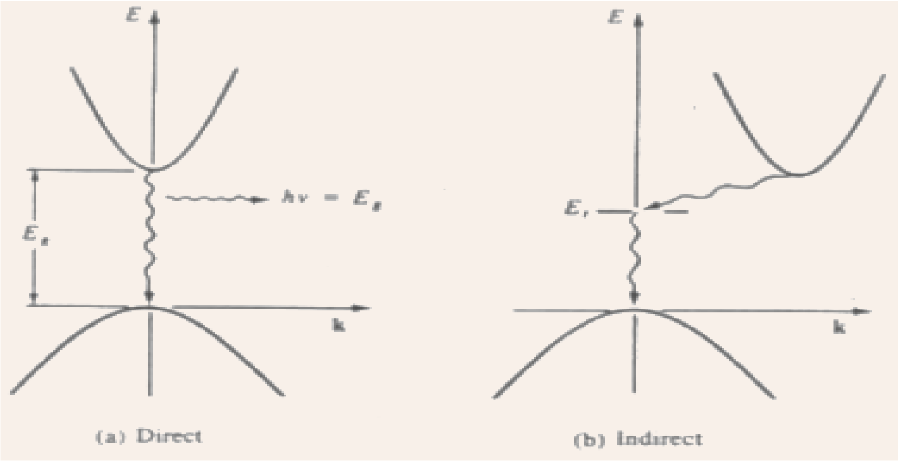
Debye length (LD), is a characteristic length for semiconductors gives an idea of the limit of the potential change in response to an abrupt change in the doping profile and is defined as,

$$L_D \equiv \sqrt{\frac{\varepsilon_s kT}{q^2 N}} = \sqrt{\frac{\varepsilon_s}{qN\beta_{th}}}.$$

Direct & indirect band-gap semiconductors:

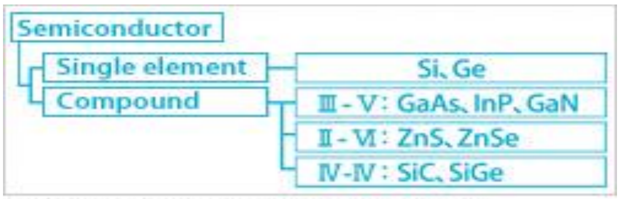
For a direct-band gap material, the minimum of the conduction band and maximum of the valance band lies at the same momentum, k, values. When an electron sitting at the bottom of the CB recombines with a hole sitting at the top of the VB, there will be no change in momentum values. Energy is conserved by means of emitting a photon, such transitions are called as radiative transitions.

Direct-band gap Semiconductor (e.g. GaAs, InP, AlGaAs)



For an indirect-band gap material the minimum of the CB and maximum of the VB lie at different k-values. When an e- and hole recombine in an indirect-band gap s/c, phonons must be involved to conserve momentum. Indirect-band gap Semiconductor (e.g. Si and Ge).

The transition that involves in an indirect band gap Semiconductor phonons without producing photons are called nonradiative (radiationless) transitions,and result in inefficient photon producing. So in order to have efficient LED's and LASER's, one should choose materials having direct band gaps such as compound Semiconductor of GaAs, AlGaAs, etc. compound semiconductor: semiconductor composed composed of two or more elements is called a compound semiconductor. Typical examples of compound semiconductors include gallium arsenide (GaAs), gallium nitride (GaN), indium phosphide (InP), zinc selenide (ZnSe), and silicon carbide (SiC).



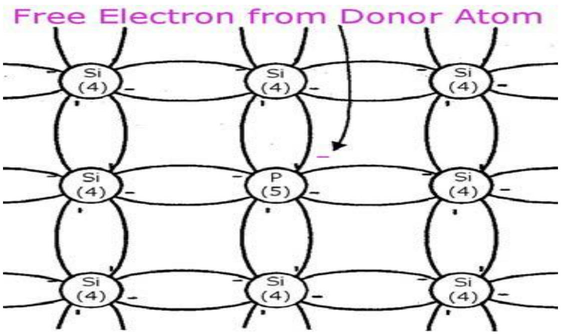
▲ Single element semiconductor and compound semiconductor

Charge Carriers in Semiconductors: Intrinsic & extrinsic semiconductor:

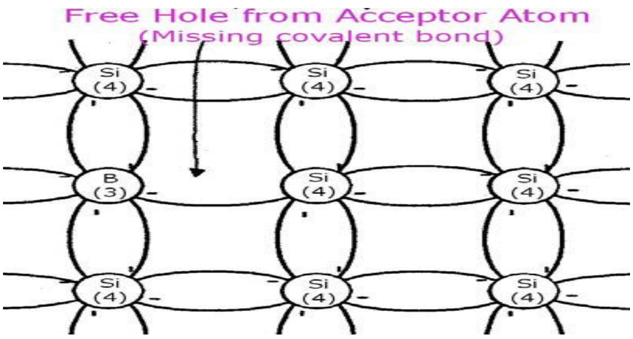
For intrinsic semiconductors at finite temperatures,thermal agitation occurs which results in continuous excitation of electrons from the valence band to the conduction band, and leaving an equal number of holes in the valence band. This process is balanced by recombination of the electrons in the conduction band with holes in the valence band. At steady state, the net result is n = p = ni , where ni is the intrinsic carrier density.

Doping is a method of selectively increasing carrier concentration, by addition of selected impurities to an intrinsic semiconductor. This is called an extrinsic semiconductor. In any semiconductor at equilibrium, the law of mass action should be satisfied i.e.

$$np = n^2_i$$



n-type Silicon



p-type Silicon

Effect of temperature and energy gap on intrinsic concentration, effect of temperature on extrinsic semiconductor

An intrinsic semiconductor several factors come to mind:

1. It is extremely pure, containing an insignificant amount of impurities.
2. The properties of the material depend only on the element(s) the semiconductor is made of.
3. For every electron created, a hole is created also, $n_o = p_o = n_i$.

For an electron-hole pair to be created in an intrinsic semiconductor, a bond must be broken in the lattice, and this requires energy. An electron in the valence band must gain enough energy to jump to the conduction band and leave a hole behind. ni represents the intrinsic carrier concentration, or we can see it as the number of bonds broken in an intrinsic semiconductor.

As the temperature is increased, the number of broken bonds (carriers) increases because there is more thermal energy available so more and more electrons gain enough energy to break free. Each electron that makes it to the conduction band leaves behind a hole in the valence band and there is an increase in both the electron and hole concentration. As the temperature is decreased, electrons do not receive enough energy to break a bond and remain in the valence band. If electrons are in the conduction band they will quickly lose energy and fall back to the valence band, annihilating a hole. Therefore, lowering the temperature causes a decrease in the intrinsic carrier concentration, while raising the temperature causes an increase in intrinsic carrier concentration.

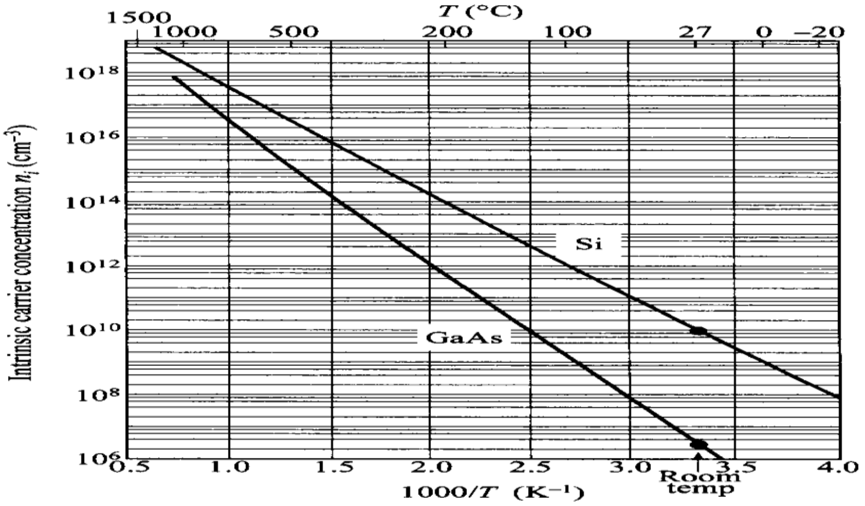


Fig.Intrinsic carrier concentrations of Si **and** GaAs as a function of reciprocal temperature

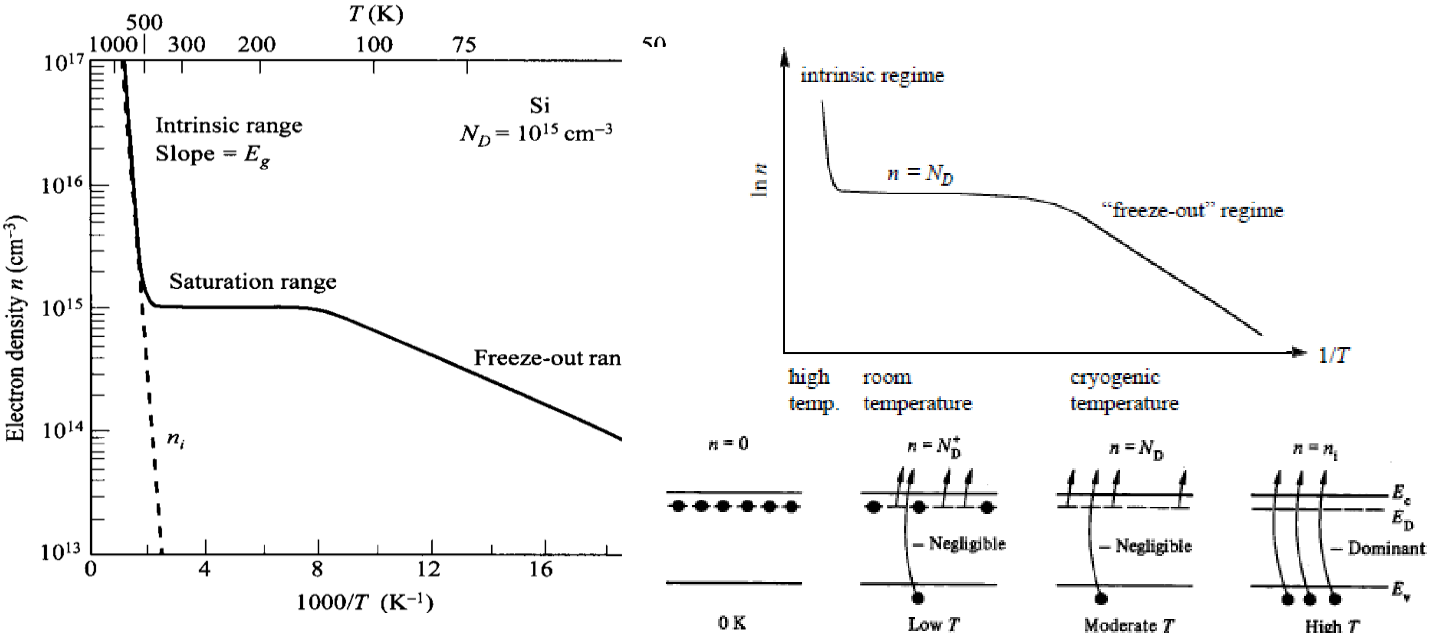


Fig. Electron density as a function of temperature for a Si sample with donor impurity concentration of 1015 cm-3

Basic concept on optical absorption ,photoluminescence, carrier life time , carrier generation and recombination

Whenever the thermal-equilibrium condition of a semiconductor system is disturbed(i.e., pn ≠ n_i2), processes exist to restore the system to equilibrium (i.e., pn ≠ n_i^2).These processes are recombination when pn > n' and thermal generation when pn < n_i^2 .

Figure… illustrates the band-to-band electron-hole recombination. The
Energy of an electron in transition from the conduction band to the valence band is conserved by emission of a photon (radiative process) or by transfer of the energy to another free electron or hole (Auger process). The former process is the inverse of direct optical absorption, and the latter is the inverse of impact ionization.

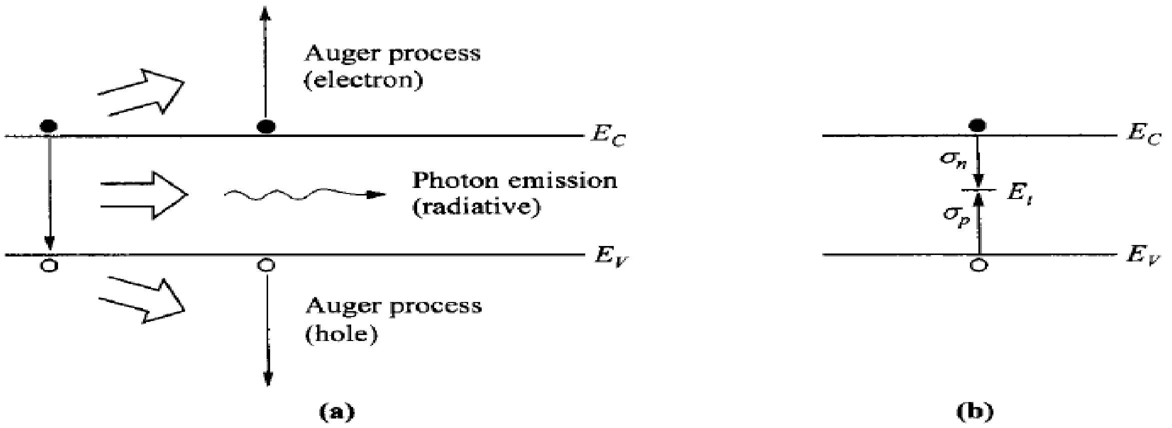


Fig. a),b) Recombination processes (the reverse are generation processes). (a) Band-to-band recombination. Energy is exchanged to a radiative or Auger process. (b) Recombination through single-level traps (nonradiative).

For this type of transition, the recombination rate is proportional to the product of electron and hole concentrations given by
Re = Rec pn.

The term Rec called the recombination coeficient, is related to the thermal generation
rate Gth, by

$$R_{ec} = \frac{G_{th}}{n_i^2} .$$

Rec, is a function of temperature and is also dependent on the band structure of the semiconductor.

$$U = R_e - G_{th} = R_{ec}(pn - n_i^2)$$
$$\approx R_{ec}\Delta p N_D \equiv \frac{\Delta p}{\tau_p}$$

where the carrier lifetime for holes

$$\tau_p = \frac{1}{R_{ec}N_D} ,$$

and in p-type material,

$$\tau_n = \frac{1}{R_{ec}N_A} .$$

However, in indirect-bandgap semiconductors such as Si and Ge, the dominant transitions are indirect recombinationlgeneration via bulk traps, of density Nt and energy E, present within the bandgap fig b

Continuity equation:
continuity equations deal with time-dependent phenomena such as low-level injection, generation and recombination. Qualitatively, the net change of carrier concentration is the difference between generation and recombination, plus the net current flowing in and out of the region of interest. The continuity equations are:

$$\frac{\partial n}{\partial t} = G_n - U_n + \frac{1}{q}\nabla \cdot J_n,$$

$$\frac{\partial p}{\partial t} = G_p - U_p - \frac{1}{q}\nabla \cdot J_p$$

where Gn, and Gp are the electron and hole generation rate (Cm-3—S-1) respectively, caused by external influences such as the optical excitation with photons or impact ionization under large electric fields.

Degeneracy and non-degeneracy of semiconductor:
If dopants added at much higher concentrations, dopant atoms come much closer to each other and it is no longer valid to assume the donor levels as atom like. If the inter-atomic distance is closer (typically < 10nm) then the atomic levels turn into bands. This leads to significant changes in the crystal structure as well as the physical properties. Another very important effect is, highly doped semiconductors come to freeze-out at much lower temperatures, meaning the freeze-out region is almost eliminated.Such highly doped semiconductors are called **Degenerate** semiconductors.

It is also to be noted that the donors (or acceptors) energy levels are assumed as atom like. Such assumptions are limited up to a certain level of dopant concentration and such extrinsic semiconductors are called *non-degenerate* semiconductors. In nondegenerate semiconductors, the doping concentrations are smaller than Nc(the effective density of states in the conduction band) and the Fermi levels are more than several kT below EC.

Non degenerate semiconductors are those which:
-are lightly doped
-have less value of electron and hole concentration
-violate Pauli's exclusion principle

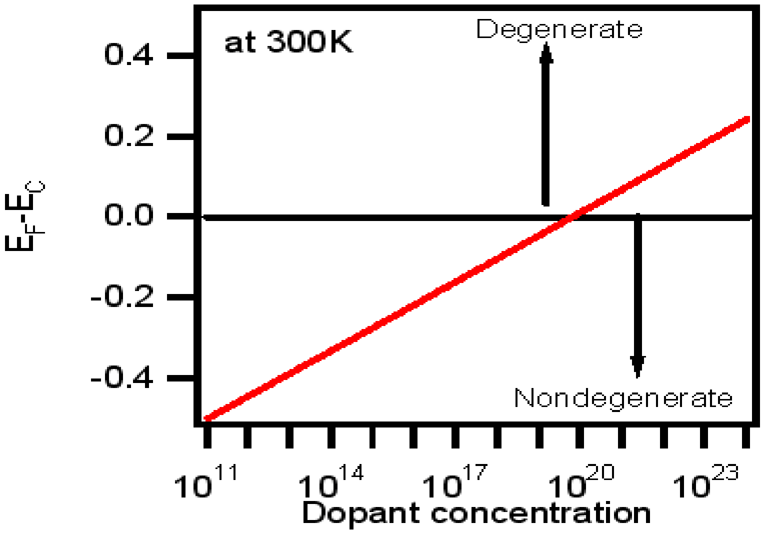


Fig. Effect of dopant (donor) concentration on Fermi level position, with respect to conduction band in silicon.

Carrier Concentration and Fermi Level, Fermi Level shift with doping & temperature, invariance of Fermi level at equilibrium:

We first consider the intrinsic case without impurities added to the semiconductor. The number of electrons (occupied conduction-band levels) is given by the total number of states N(E) multiplied by the occupancy F(E), integrated over the conduction band,

$$n = \int_{E_C}^{\infty} N(E)F(E)dE$$

The density of states N(E) can be approximated by the density near the bottom of the conduction band for low-enough carrier densities and temperature.The occupancy is a strong function of temperature and energy, and is represented by the Fermi-Dirac distribution Function

$$F(E) = \frac{1}{1 + \exp[(E - E_F)/kT]}$$

Where E_F is the Fermi energy level which can be determined from the charge neutrality condition

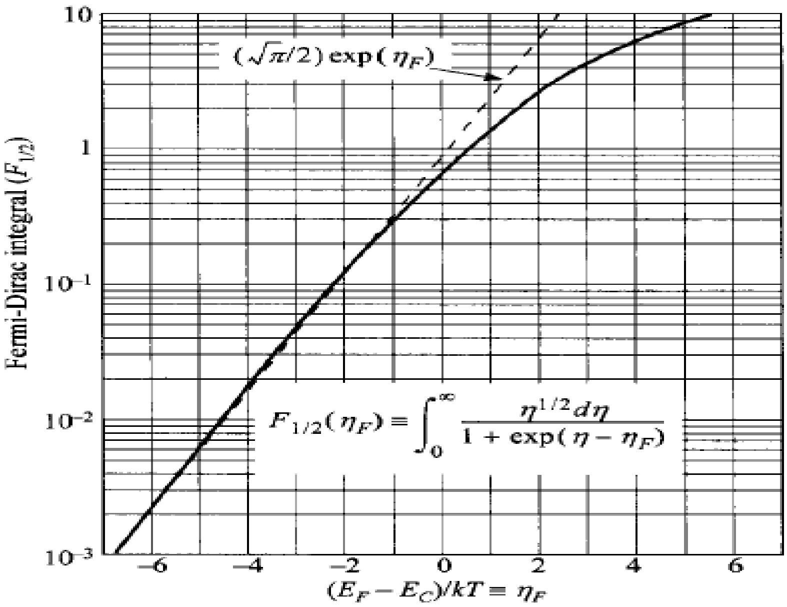


Fig Fermi-Dirac integral $F_{1/2}$, as a function of Fermi energy. Dashed line is approximation of Boltzmann statistics.

As shown in the fig for degenerate levels where n- or p-concentrations are near or beyond the effective density of states (Ncor Nv), the value of Fermi-Dirac integral has to be used instead of the simplified Boltzmann statistics.

Intrinsic carrier concentration:

For intrinsic semiconductors at finite temperatures, thermal agitation occurs which results in continuous excitation of electrons from the valence band to the conduction band, and leaving an equal number of holes in the valence band. This process is balanced by recombination of the electrons in the conduction band with holes in the valence band. At steady state, the net result is n = p = ni , where ni is the intrinsic carrier density.

$$E_F = E_i = \frac{E_C + E_V}{2} + \frac{kT}{2}\ln\left(\frac{N_V}{N_C}\right)$$

Hence the Fermi level Ei of an intrinsic semiconductor generally lies very close to, but not exactly at, the middle of the bandgap. The intrinsic carrier density ni can be obtained from

$$n_i = N_C\exp\left(-\frac{E_C - E_i}{kT}\right) = N_V\exp\left(-\frac{E_i - E_V}{kT}\right) = \sqrt{N_C N_V}\exp\left(-\frac{E_g}{2kT}\right)$$

Figure 9 shows the temperature dependence of ni for Si and GaAs. As expected, the larger the bandgap is, the smaller the intrinsic carrier density will be. It also follows that for nondegenerate semiconductors, the product of the majority and minority carrier concentrations is fixed to be.

$$\begin{aligned} pn &= N_C N_V \exp\left(-\frac{E_g}{kT}\right) \\ &= n_i^2 \quad , \end{aligned}$$

which is known as the mass-action law. But for degenerate semiconductors, pn < ni2; the alternate equations for

n-type materials;

$$n = n_i \exp\left(\frac{E_F - E_i}{kT}\right) \quad \text{or} \quad E_F - E_i = kT \ln\left(\frac{n}{n_i}\right)$$

and for *p*-type materials;

$$p = n_i \exp\left(\frac{E_i - E_F}{kT}\right) \quad \text{or} \quad E_i - E_F = kT \ln\left(\frac{p}{n_i}\right)$$

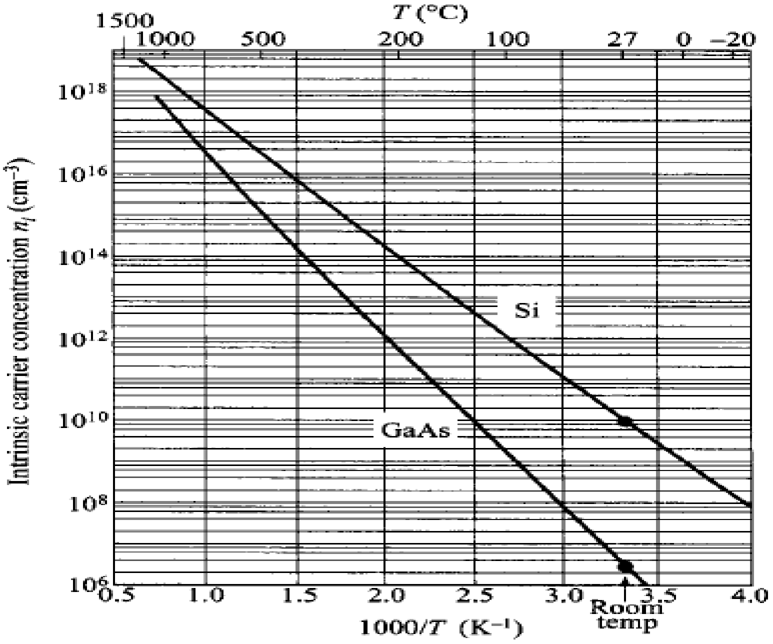
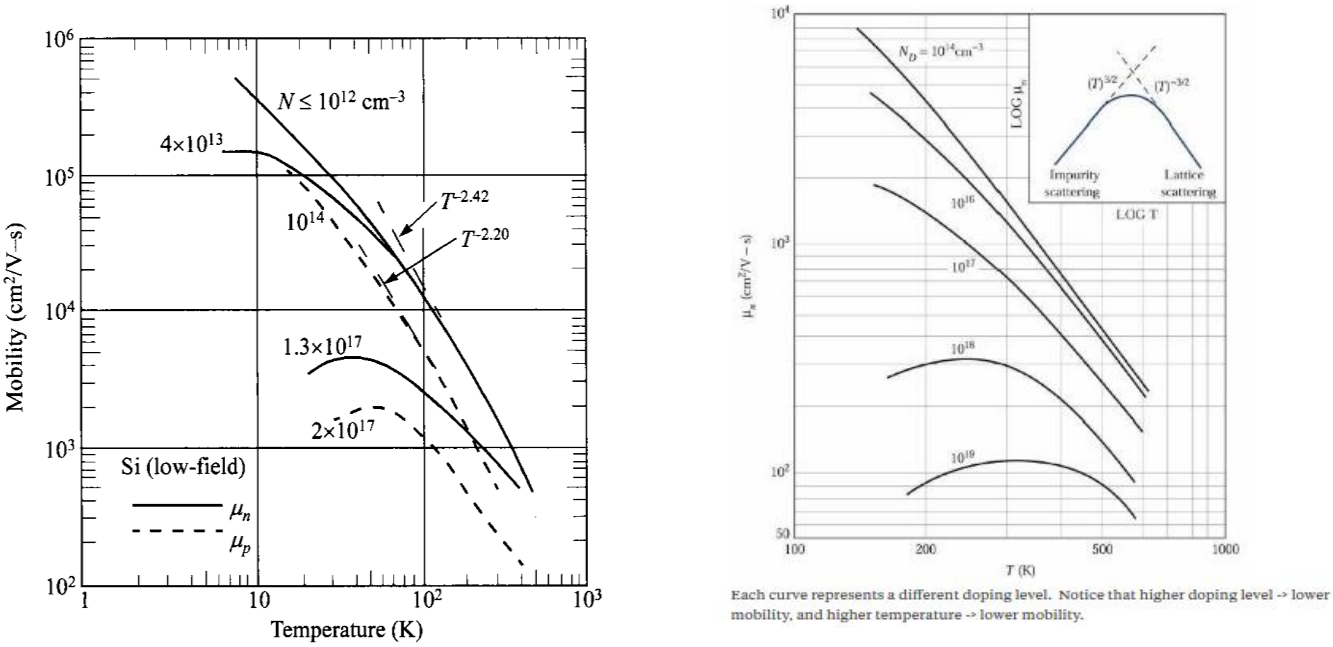


Fig.Intrinsic carrier concentrations of Si and GaAs as a function of reciprocal temperature.

Non-equilibrium condition: Effect of temperature and doping concentration on mobility:

Conductivity of a material is determined by two factors: the concentration of free carriers available to conduct current and their mobility (or freedom to move). In a semiconductor, both mobility and carrier concentration are temperature dependent.

Both increasing temperature and increasing doping levels have the tendency to reduce electron and hole mobility. For temperature, the increased temperature increases the number of phonons, which increases the probability that an electron will be scattered by a phonon. For doping levels, each dopant atom is a defect site that an electron can scatter from.



Each curve represents a different doping level. Notice that higher doping level -> lower mobility, and higher temperature -> lower mobility.

Fig. Mobility of electrons and holes in Si as a function of temperature

Drift & diffusion of carriers:

The drift current is due to the motion of charge carriers due to the force exerted on them by an electric field.

Whenever there exists a gradient of carrier concentration, a process of diffusion occurs by which the carriers migrate from the region of high concentration toward the region of low concentration, to drive the system toward a state of uniformity. This flow or flux of carriers, taking electrons as an example, is governed by the Fick's law,

The most-common current conduction consists of the drift component, caused by the electric field, and the diffusion component, caused by the carrier-concentration gradient. The current-density equations are:

$$J_n = q\mu_n n\mathscr{E} + qD_n\nabla n$$
$$J_p = q\mu_p p\mathscr{E} - qD_p\nabla p$$
$$J_{\text{cond}} = J_n + J_p ,$$

where Jn, and Jp are the electron and hole current densities, respectively.

Hall Effect and piezo electric effect:

The Hall effect is the generation of a Hall voltage VH, when a piece of semiconductor is biased with a current and placed under a magnetic field that is orthogonal to the current flow. The generated Hall voltage, assuming a Hall factor rH = 1 and a p-type semiconductor, is given by

$$V_H = R_H W J_x \mathscr{B} = W\mathscr{E}_x \mu \mathscr{B}$$

The Hall effect is used in common practice to measure certain properties of semiconductors: namely, the carrier concentration, mobility, and the type (n or p). It is an important analytical tool since a simple conductance measurement can only give the product of concentration and mobility, and the type remains unknown.

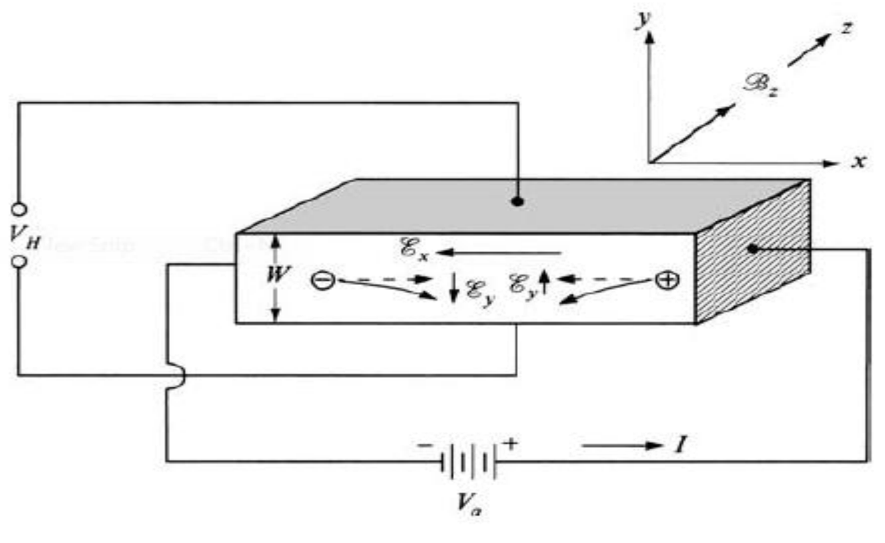


Fig.Basic setup to measure carrier concentration using the Hall effect

Piezoelectric Effect is the ability of certain materials to generate an electric charge in response to applied mechanical stress. The word Piezoelectric is derived from the Greek piezein, which means to squeeze or press, and piezo, which is Greek for “push”.

One of the unique characteristics of the piezoelectric effect is that it is reversible, meaning that materials exhibiting the direct piezoelectric effect (the generation of electricity when stress is applied) also exhibit the converse piezoelectric effect (the generation of stress when an electric field is applied).

Common piezoelectric materials are quartz, LiNbO,, ZnO, BaTiO,, LiTaO,, and lead zirconate titanates. These materials are also good insulators. Less common piezoelectric materials are semiconductors such as CdS, CdSe, CdTe, and GaAs. A prerequisite of the piezoelectric effect is some degree of lattice order, so crystal or polycrystal structures are required.

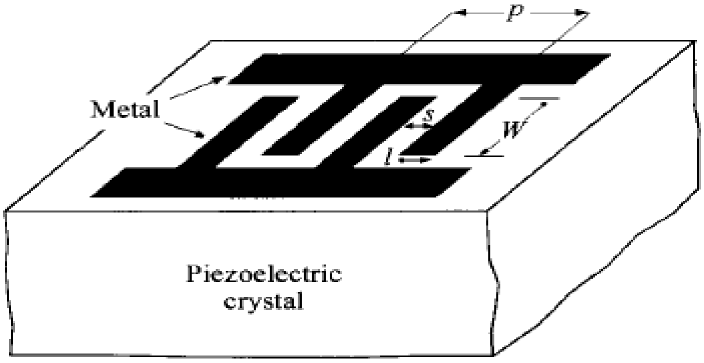


Fig.Interdigital transducer on bulk piezoelectric substrate.

Module II: Junction Physics in Semiconductor Devices:

2.1 PN Junction Diode :

A *PN Junction Diode* is one of the simplest semiconductor devices around, and which has the characteristic of passing current in only one direction only. However, unlike a resistor, a diode does not behave linearly with respect to the applied voltage as the diode has an exponential current-voltage (I-V) relationship and therefore we can not described its operation by simply using an equation such as Ohm's law.

If a suitable positive voltage (forward bias) is applied between the two ends of the PN junction, it can supply free electrons and holes with the extra energy they require to cross the junction as the width of the depletion layer around the PN junction is decreased.

By applying a negative voltage (reverse bias) results in the free charges being pulled away from the junction resulting in the depletion layer width being increased. This has the effect of increasing or decreasing the effective resistance of the junction itself allowing or blocking current flow through the diode.

Then the depletion layer widens with an increase in the application of a reverse voltage and narrows with an increase in the application of a forward voltage. This is due to the differences in the electrical properties on the two sides of the PN junction resulting in physical changes taking place. One of the results produces rectification as seen in the PN junction diodes static I-V (current-voltage) characteristics. Rectification is shown by an asymmetrical current flow when the polarity of bias voltage is altered as shown below.

2.2 Energy band diagram:

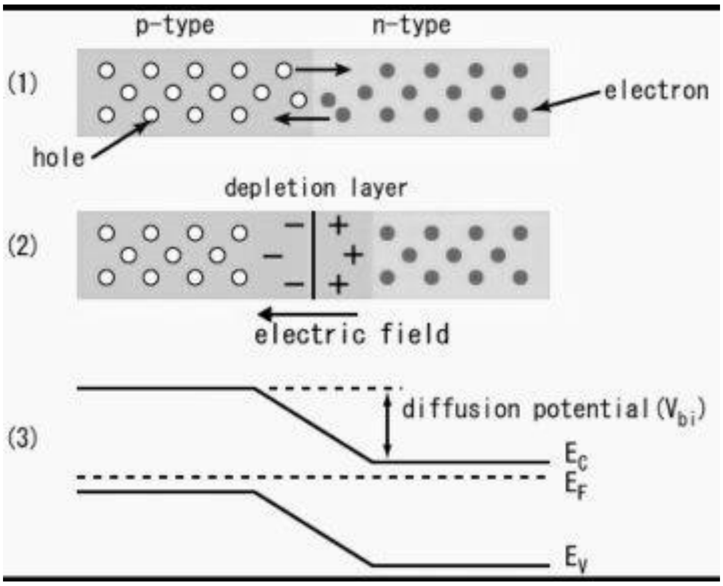


Fig. 2.1 Energy band diagra

2.3 Junction Diode Symbol and Plotting of junction voltage (Static I-V Characteristics)

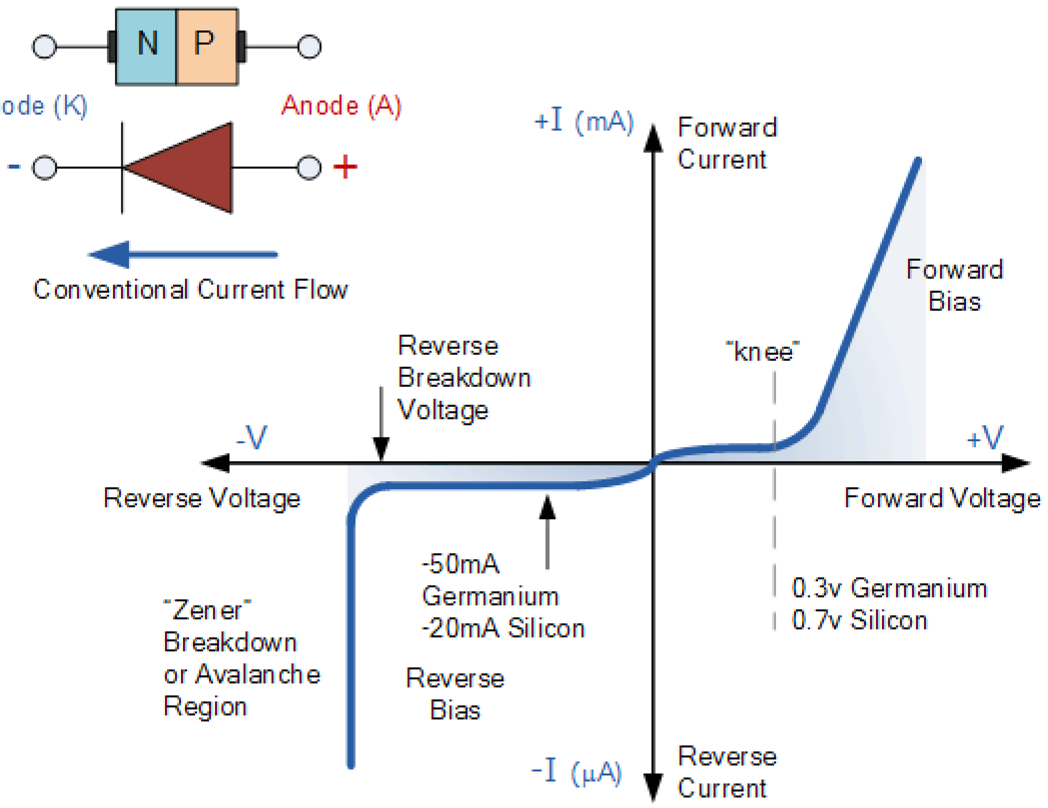


Fig. 2.2 Plotting of junction voltage

But before we can use the PN junction as a practical device or as a rectifying device we need to firstly bias the junction, ie connect a voltage potential across it. On the voltage axis above, "Reverse Bias" refers to an external voltage potential which increases the potential barrier. An external voltage which decreases the potential barrier is said to act in the "Forward Bias" direction.

There are two operating regions and three possible "biasing" conditions for the standard Junction Diode and these are:

Zero Bias – No external voltage potential is applied to the PN junction diode.

Reverse Bias – The voltage potential is connected negative, (-ve) to the P-type material and positive, (+ve) to the N-type material across the diode which has the effect of Increasing the PN junction diode's width.

Forward Bias – The voltage potential is connected positive, (+ve) to the P-type material and negative, (-ve) to the N-type material across the diode which has the effect of Decreasing the PN junction diodes width.

2.4 Zero Biased PN Junction Diode

When a diode is connected in a Zero Bias condition, no external potential energy is applied to the PN junction. However if the diodes terminals are shorted together, a few holes (majority carriers) in the P-type material with enough energy to overcome the potential barrier will move across the junction against this barrier potential. This is known as the "Forward Current" and is referenced as IF

Likewise, holes generated in the N-type material (minority carriers), find this situation favourable and move across the junction in the opposite direction. This is known as the "Reverse Current" and is referenced as IR. This transfer of electrons and holes back and forth across the PN junction is known as diffusion, as shown below.

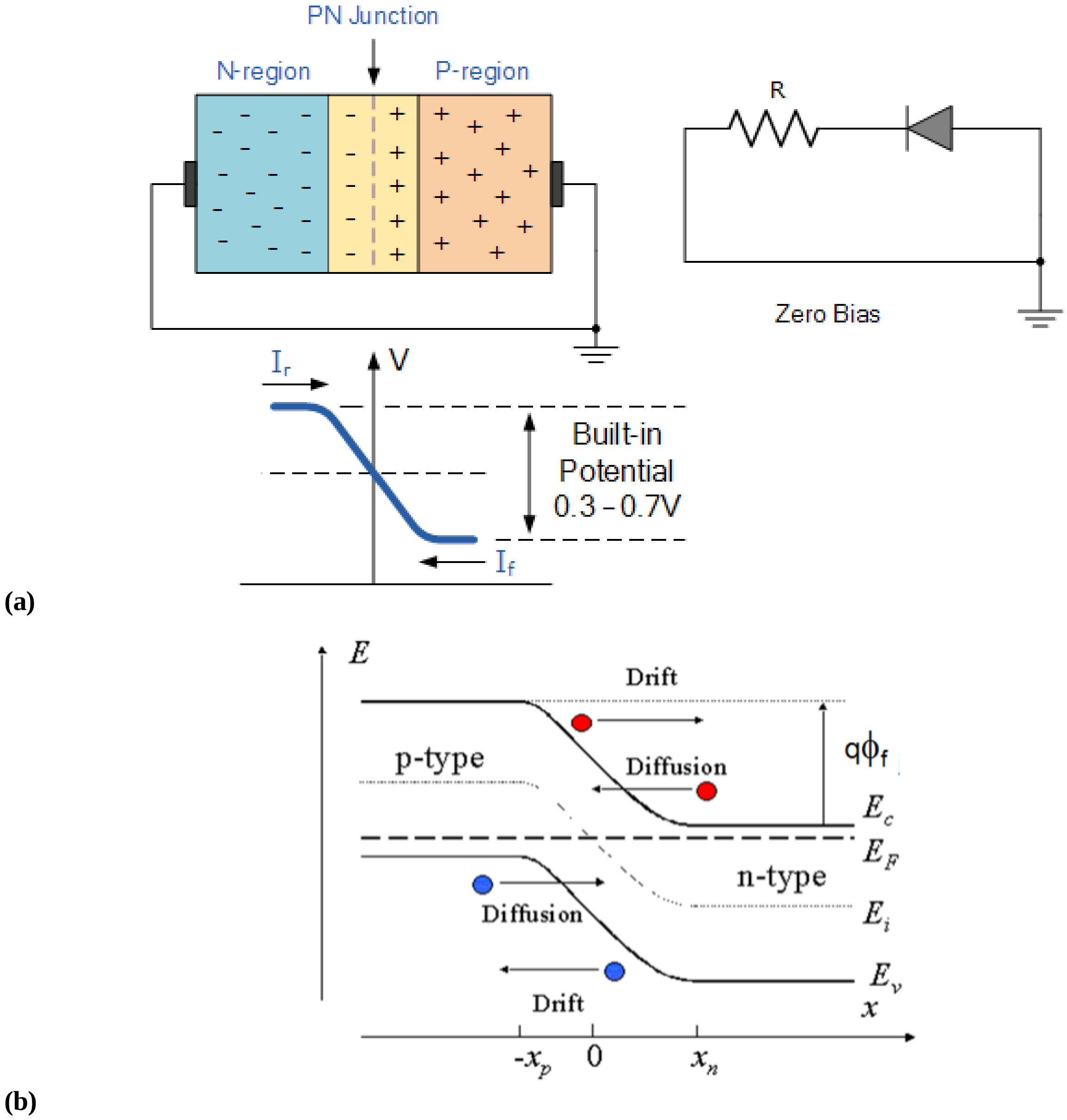


Fig. 2.3 (a) Schematic (b) detailed energy band diagram under zero bias

The potential barrier that now exists discourages the diffusion of any more majority carriers across the junction. However, the potential barrier helps minority carriers (few free electrons in the P-region and few holes in the N-region) to drift across the junction.

Then an "Equilibrium" or balance will be established when the majority carriers are equal and both moving in opposite directions, so that the net result is zero current flowing in the circuit. When this occurs the junction is said to be in a state of "Dynamic Equilibrium".

The minority carriers are constantly generated due to thermal energy so this state of equilibrium can be broken by raising the temperature of the PN junction causing an increase in the generation of minority carriers, thereby resulting in an increase in leakage current but an electric current cannot flow since no circuit has been connected to the PN junction.

2.5 Forward Biased PN Junction Diode

When a diode is connected in a Forward Bias condition, a negative voltage is applied to the N-type material and a positive voltage is applied to the P-type material. If this external voltage becomes greater than the value of the potential barrier, approx. 0.7 volts for silicon and 0.3 volts for germanium, the potential barriers opposition will be overcome and current will start to flow.

This is because the negative voltage pushes or repels electrons towards the junction giving them the energy to cross over and combine with the holes being pushed in the opposite direction towards the junction by the positive voltage. This results in a characteristics curve of zero current flowing up to this voltage point, called the "knee" on the static curves and then a high current flow through the diode with little increase in the external voltage as shown below.

2.5.1 Reduction in the Depletion Layer due to Forward Bias current components in forward and reverse biased junction

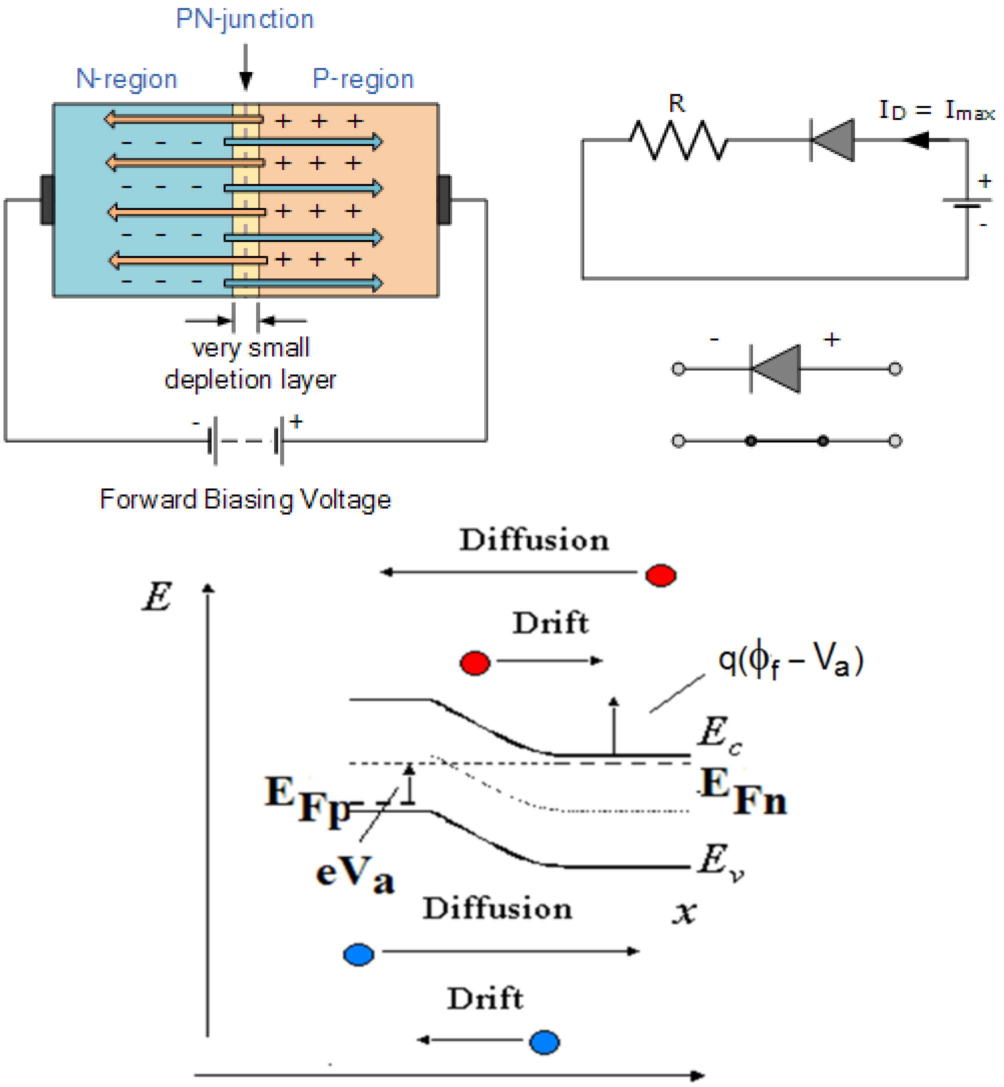


Fig. 2.4 (a) Schematic (b) detailed energy band diagram under positive bias

This condition represents the low resistance path through the PN junction allowing very large currents to flow through the diode with only a small increase in bias voltage. The actual potential difference across the junction or diode is kept constant by the action of the depletion layer at approximately 0.3v for germanium and approximately 0.7v for silicon junction diodes.

Since the diode can conduct “infinite” current above this knee point as it effectively becomes a short circuit, therefore resistors are used in series with the diode to limit its current flow. Exceeding its maximum forward current specification causes the device to dissipate more power in the form of heat than it was designed for resulting in a very quick failure of the device.

2.6 Reverse Biased PN Junction Diode

2.6.1 Creation of depletion region

When a diode is connected in a Reverse Bias condition, a positive voltage is applied to the N-type material and a negative voltage is applied to the P-type material.

The positive voltage applied to the N-type material attracts electrons towards the positive electrode and away from the junction, while the holes in the P-type end are also attracted away from the junction towards the negative electrode.

The net result is that the depletion layer grows wider due to a lack of electrons and holes and presents a high impedance path, almost an insulator. The result is that a high potential barrier is created thus preventing current from flowing through the semiconductor material and this potential is called built in potential.

2.6.1.1 Built in potential:

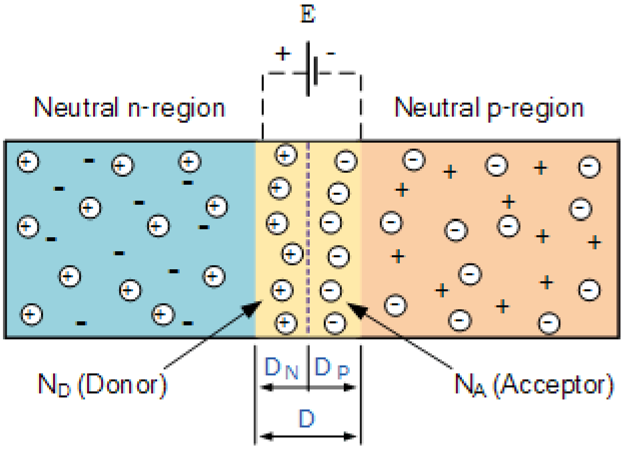


Fig.2.5 Built in potential

As the N-type material has lost electrons and the P-type has lost holes, the N-type material has become positive with respect to the P-type. Then the presence of impurity ions on both sides of the junction cause an electric field to be established across this region with the N-side at a positive voltage relative to the P-side. The problem now is that a free charge requires some extra energy to overcome the barrier that now exists for it to be able to cross the depletion region junction.

This electric field created by the diffusion process has created a "built-in potential difference" across the junction with an open-circuit (zero bias) potential of:

$$E_o = V_T \ln\left(\frac{N_D.N_A}{n_i^2}\right)$$

Where: Eo is the zero bias junction voltage, VT the thermal voltage of 26mV at room temperature, ND and NA are the impurity concentrations and ni is the intrinsic concentration.

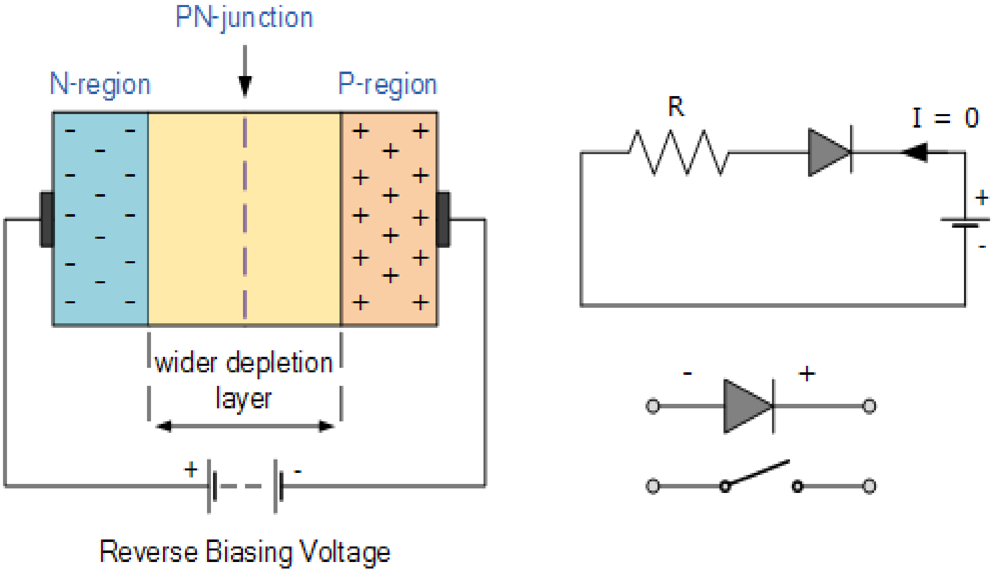
A suitable positive voltage (forward bias) applied between the two ends of the PN junction can supply the free electrons and holes with the extra energy. The external voltage required to overcome this potential barrier that now exists is very much dependent upon the type of semiconductor material used and its actual temperature.

Typically at room temperature the voltage across the depletion layer for silicon is about 0.6 – 0.7 volts and for germanium is about 0.3 – 0.35 volts. This potential barrier will always exist even if the device is not connected to any external power source, as seen in diodes.

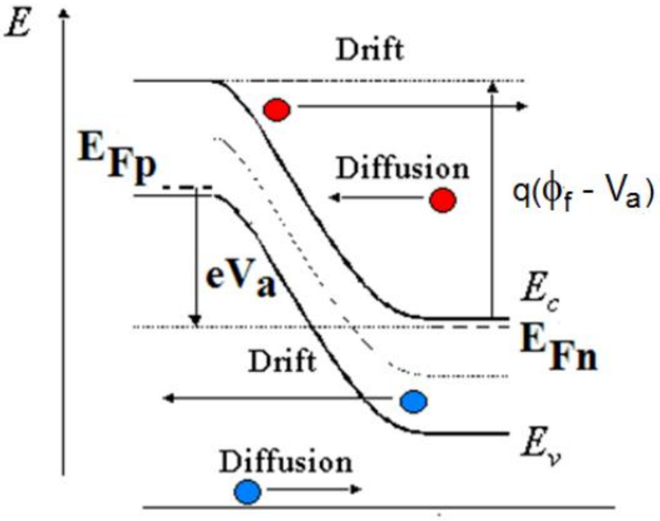
The significance of this built-in potential across the junction, is that it opposes both the flow of holes and electrons across the junction and is why it is called the potential barrier. In practice, a PN junction is formed within a single crystal of material rather than just simply joining or fusing together two separate pieces.

The result of this process is that the PN junction has rectifying current–voltage (IV or I–V) characteristics. Electrical contacts are fused onto either side of the semiconductor to enable an electrical connection to be made to an external circuit. The resulting electronic device that has been made is commonly called a PN junction Diode or simply Signal Diode.

2.6.2 Increase in the Depletion Layer due to Reverse Bias



(a)



(b)

Fig. 2.6 (a) Schematic (b) detailed energy band diagram under negative bias

This condition represents a high resistance value to the PN junction and practically zero current flows through the junction diode with an increase in bias voltage. However, a very small leakage current does flow through the junction which can be measured in micro-amperes, (µA).

One final point, if the reverse bias voltage Vr applied to the diode is increased to a sufficiently high enough value, it will cause the diode's PN junction to overheat and fail due to the avalanche effect around the junction. This may cause the diode to become shorted and will result in the flow of maximum circuit current, and this shown as a step downward slope in the reverse static characteristics curve below.

2.7 Current components in forward and reverse biased junction

2.7.1 Forward Characteristics Curve for a Junction Diode

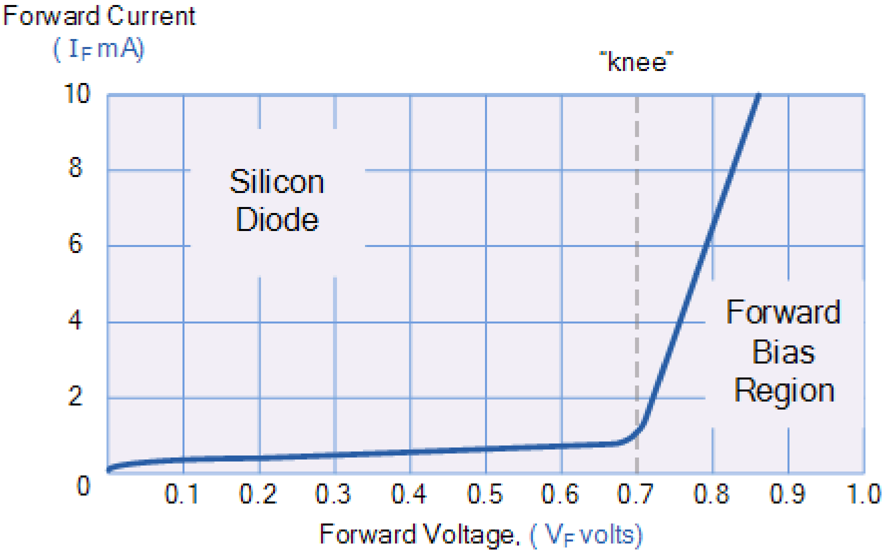


Fig. 2.7 Forward Characteristics Curve

The application of a forward biasing voltage on the junction diode results in the depletion layer becoming very thin and narrow which represents a low impedance path through the junction thereby allowing high currents to flow. The point at which this sudden increase in current takes place is represented on the static I-V characteristics curve above as the "knee" point.

2.7.2 Reverse Characteristics Curve for a Junction Diode

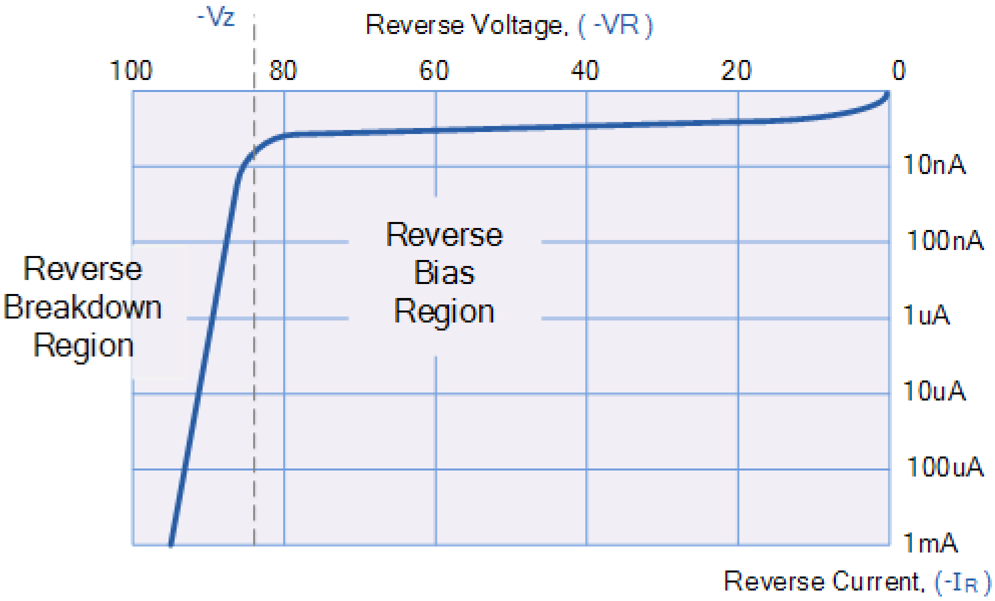


Fig. 2.8 Reverse Characteristics Curve

Sometimes this avalanche effect has practical applications in voltage stabilising circuits where a series limiting resistor is used with the diode to limit this reverse breakdown current to a preset maximum value thereby producing a fixed voltage output across the diode. These types of diodes are commonly known as Zener Diodes and are discussed in a later tutorial.

FAQ

How electric field strength increases with the reverse bias?
Total voltage across the pn junction will be V+Vx. The electrons at n-side will get pulled from junction region to the terminal region of n-side and similarly the holes at p-side junction will get pulled towards the terminal region of p-side. This results in increasing the depletion region width from its initial length, say ‘W’ to some ‘W+x’. As width of depletion region increases, it results in increasing the electric field strength.

How reverse saturation current occurs and why it exists ?
The reverse saturation current is the negligibly small current (in the range of micro amperes) shown in graph, from 0 volts to break down voltage. It remains almost constant (negligible increase do exist) in the range of 0 volts to reverse breakdown voltage. How it occurs ? We know, as electrons and holes are pulled away from junction, they dont get diffused each other across the junction. So the net “diffusion current” is zero! What remains is the drift due to electric field. This reverse saturation current is the result of drifting of charge carriers from the junction region to terminal region. This drift is caused by the electric field generated by depletion region.

What happens at reverse breakdown ?
At breakdown voltage, the current through diode shoots rapidly. Even for a small change in applied voltage, there is a high increase in net current through the diode. For each pn junction diode, there will be a maximum net current that it can withstand. If the reverse current exceeds this maximum rating, the diode will get damaged.
2.8 Diode Equation
The voltage developed across a p-n junction caused by –
the diffusion of electrons from the n-side of the junction into the p side and
the diffusion of holes from the p side of the junction into the n side
Drift currents only flow when there is an electric field present.
Diffusion currents only flow when there is a concentration difference for either the electrons or holes (or both).

$$I_n^{drift} = qA\mu_n nE$$

$$I_p^{drift} = qA\mu_p pE$$

$$I^{drift} = Aq(\mu_n n + \mu_p p)E$$

$$I_n^{diff} = qAD_n \nabla n = qAD_n \frac{dn}{dx}$$

$$I_p^{diff} = -qAD_p \nabla p = -qAD_p \frac{dp}{dx}$$

$$I^{diff} = I_n^{diff} + I_p^{diff} = qA(D_n \nabla n - D_p \nabla p$$

$$I^T = I^{diff} + I^{drift}$$

Ideal Diodes

The diode equation gives an expression for the current through a diode as a function of voltage. The *Ideal Diode Law*, expressed as:

$$I = I_0\left(e^{\frac{qV}{kT}} - 1\right)$$

where:
I = the net current flowing through the diode;
I_0 = "dark saturation current", the diode leakage current density in the absence of light;
V = applied voltage across the terminals of the diode;
q = absolute value of electron charge;
k = Boltzmann's constant; and
T = absolute temperature (K).

The "dark saturation current" (I_0) is an extremely important parameter which differentiates one diode from another. I_0 is a measure of the recombination in a device. A diode with a larger recombination will have a larger I_0. An excellent discussion of the recombination parameter is in [1]

Note that:

- I_0 increases as *T* increases; and
- I_0 decreases as material quality increases.

At 300K, *kT/q* = 25.85 mV, the "thermal voltage".

Lp = diffusion length of holes (cm)
Dp = diffusion constant (cm2/s) L p Lp = average carrier life time (s)

Firstly, we find Δp where Δp is the minority carrier concentration at the edge of the depletion region (DP) we know that the built-in voltage is given by V $_{b i}$ = k T/ q ln (N $_A$ N $_D$ n $_i$ ^{2}) V$_{bi}$ =kT/q ln⁡(N$_A$N$_D$n$_i^2$)
Applying the law of mass action n $_i$ ^{2} = n $_{n o}$ × p $_{n o}$
we get V $_{b i}$ = k T/ q ln (n $_{n o}$ p $_{n o}$ n i 2) Vbi=kTqln⁡(nnopnoni2) Rearranging p p o = p n o exp (q V b i k T) ppo=pnoexp⁡(qVbikT) => eqn.1 For non-equilibrium situation, i.e. when there's forward bias voltage V f Vf p p (0) = p n (0) exp (q (V b i − V f) k T) pp(0)=pn(0)exp⁡(q(Vbi−Vf)kT) => eqn.2 or p p o = p n o exp (q V k T) ppo=pnoexp⁡(qVkT) where V = V b i − V f V=Vbi−Vf assuming low injection level, i.e. p p ≈ p p o pp≈ppo eqn.1/eqn.2 by doing this we get Δ p Δp therefore Δ p = p n o exp (q V k T − 1) Δp=pnoexp⁡(qVkT−1) Using the continuity equation, we get an expression for the current density J p (x) = q D p L p δ p (x) Jp(x)=qDpLpδp(x) since Δ p = δ p (x = 0) Δp=δp(x=0) so δ p (x = 0) = p n o exp (q V k T − 1) δp(x=0)=pnoexp⁡(qVkT−1) J p (x) = q D p p n o L p exp (q V k T − 1) Jp(x)=qDppnoLpexp⁡(qVkT−1) Doing the same with electrons, we get a similar expression, hence J t o t a l = J p + J n = J s exp (q V k T − 1) Jtotal=Jp+Jn=Jsexp⁡(qVkT−1) This is the standard way of expressing the diode equation. However, if we multiply the above expression by the cross-section area, we get the current I.

Reference https://www.physicsforums.com/threads/simple-derivation-of-diode-equation.307717/

Varactor Diode

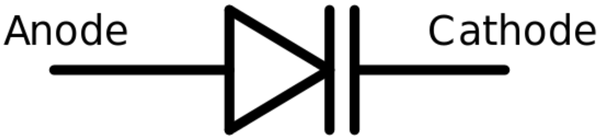


Fig. 2.9 Symbol of Varactor Diode

Varactor Diode is a reverse biased p-n junction diode, whose capacitance can be varied electrically. As a result these diodes are also referred to as varicaps, tuning diodes, voltage variable capacitor diodes, parametric diodes and variable capacitor diodes. It is well known that the operation of the p-n junction depends on the bias applied which can be either forward or reverse in characteristic. It is also observed that the span of the depletion region in the p-n junction decreases as the voltage increases in case of forward bias. On the other hand, the width of the depletion region is seen to increase with an increase in the applied voltage for the reverse bias scenario. Under such condition, the p-n junction can be considered to be analogous to a capacitor (Figure 1) where the p and n layers represent the two plates of the capacitor

while the depletion region acts as a dielectric separating them. Thus one can apply the formula used to compute the capacitance of a parallel plate capacitor even to the varactor diode.
Hence, mathematical expression for the capacitance of varactor diode is given by:

Where, Cj is the total capacitance of the junction. ε is the permittivity of the semiconductor material. A is the cross-sectional area of the junction. d is the width of the depletion region. Further the relationship between the capacitance and

the reverse bias voltage is given as Where, Cj is the capacitance of the varactor diode. C is the capacitance of the varactor diode when unbiased. K is the constant, often considered to be 1. Vb is the barrier potential. VR is the applied reverse voltage. m is the material dependent constant. its symbol is shown by Figure 2.9. They are used in

1. Tuning circuits to replace the old style variable capacitor tuning of FM radio
2. Small remote control circuits
3. Tank circuits of receiver or transmitter for auto-tuning as in case of TV
4. Signal modulation and demodulation.
5. Microwave frequency multipliers as a component of LC resonant circuit
6. Very low noise microwave parametric amplifiers
7. AFC circuits
8. Adjusting bridge circuits
9. Adjustable bandpass filters
10. Voltage Controlled Oscillators (VCOs)
11. RF phase shifters
12. Frequency multipliers

Zener break down principle:

Zener diode is a PN junction diode specially designed to operate in the reverse biased mode. In forward bias mode it acts as normal diode. It has a particular voltage known as Break down voltage, at which the diode breaks down while reverse biased. In the case of normal diodes, the diode damages at the break down voltage.

The basic principle of zener diode is the zener break down. When a diode is heavily doped, it's depletion region will be narrow. When a high reverse voltage is applied across the junction, there will be very strong electric field at the junction. And the electron hole pair generation takes place . Thus heavy current flows. This is known as zener breakdown.

Zener diodes are widely used as voltage references and as shunt regulators to regulate the voltage across small circuits.
The Zener Breakdown is observed in the Zener diodes having Vz less than 5V or between 5 to 8 volts. When a reverse voltage is applied to a Zener diode, it causes a very intense electric field to appear across a narrow depletion region. Such an intense electric field is strong enough to pull some of the valence electrons into the conduction band by breaking their covalent bonds .these electrons then become free electrons which are available for conduction. A large number of such free electrons will constitute a large reverse current through the Zener diode and breakdown is said to have occurred due to the Zener effect.
Dynamic resistance of Zener Diode:

Dynamic Resistance is a concept of resistance used in PN junction in Electronics. Dynamic resistance refers to the change in current in response to a change in voltage at a specific region of the VI curve.

Effect of temperature on Zener Diode:

If a Zener diode is connected to a constant current source, then at constant ambient temperature, the Zener voltage changes and approaches asymptoti- cally a final value. This voltage change is due to the power dissipated in the junction which in turn causes a rise in junction temperature. Zener diodes with a negative temperature coefficient exhibit a Zener voltage reduction, whereas those with a positive temper- ature coefficient show a Zener voltage increase on application of current. The magnitude of this voltage change due to intrinsic heat generation can be derived from the relevant curves. Because it is not practical to wait during tests until each device has r eached its thermal equilibrium, it is

common practice to measure the breakdown voltage of Zener diodes by application of a pulsating current of less than 1 sec duration. Under these conditions the junction temperature is the same as the ambient temperature. The magnitude of the test current used varies from type to type and is quoted in the relevant data sheets Therefore, designers, but especially customers carrying out acceptance tests, should allow for the fact that the Zener voltage of a device which is at thermal equi- librium will differ from that q uoted in the da

ta sheet. To arrive at an estimate of the equilibrium Zener voltage, a voltage equal to the product of Zener current and thermal differential resistance should be added to the voltage associated with the chosen current as derived from the published dynamically measured breakdown curves.

Solar Cell:

Interaction among electrons, holes, phonons, photons and other particles are required to satisfy conservation of energy and crystal momentum. A photon with an energy near a semiconductor band gap has almost zero momentum. An important process is called radiative recombination, where an electron in the conduction band annihilates a hole in the valance band, releasing the excess energy as a photon. If the electron is at the bottom of the conduction band and the hole is at the top of the valence band then in case of direct band gap semiconductor this radiative recombination is a preferred phenomenon. In case of indirect band gap semiconductor, it is not possible as it violets the conservation of crystal momentum. Though it might be possible for indirect band gap material if the process involve the absorption or emission of phonon. In that case the phonon momentum must be equals to the difference between the electron and hole momentum. The involvement of phonon makes the process slower for indirect band gap semiconductor.

Light absorption is just the reverse process of radiative recombination.

When light energy of a particular wavelength equivalent to the band gap absorption edge is incident on anindirect band gap semiconductor, it can penetrate much further than in a direct band gap semiconductor before being absorbed.

This is the key factor for photovoltaic devices. That’s why still silicon is used as solar cell substrate material though it is an indirect band gap semiconductor.

Photoelectric devices convert light energy directly into electrical energy. These devices are self-generating that means it requires no external power source to deliver the output. Photovoltaic and photoluminescence effect are just reverse to each other.

Devices that converts electricity to light exploiting the photoluminescence effect are called emitters (as it emit light) whereas devices that convert light into electricity are called photovoltaic effect. Photovoltaic devices include photoemmisive (non-solid state) devices and photodetectors (solid state devices).

Electrons are in the higher energy state on the n side whereas holes are in the lower energy band on the p side. When these electrons-holes recombine, some of this energy is given up in the form of heat and light. Generally, compound semiconductors (GaAs, GaP, GaAsP etc.) response by releasing greater percentage of energy in the form of light. If the semiconductor material is translucent , the light is emitted and the junction becomes a light source. On the other hand, when light energy of a particular wavelength equivalent to the band gap absorption edge and of particular intensity is incident on a semiconductor, the light is absorbed and breaks the covalent bond to generate electron hole pairs. These carriers cross the junction and generate electric current known as photocurrent flowing through the external circuit.

1. Physics of photovoltaic:

Solar cell is the most popular photovoltaic devices that combine optics with electronics. It converts solar energy into optical energy and also known as solar energy converter. The reaching the earth surface from the sun is basically an electromagnetic radiation which covers a spectral range of 0.2 to 0.3 micrometer.

In 1954, solar cell was first developed by Chaplin, Fuller and Pearson and since then it has shown remarkable progress in the market.

There are two parameters that ultimately used to characterize the solar cell:

i) Short circuit current: the maximum current, at zero voltage. Ideally, if V = 0, Isc = IL. Note that Isc is directly proportional to the available sunlight.

ii) Open circuit voltage: the maximum voltage, at zero current. The value of Voc increases logarithmically with increased sunlight. This characteristic makes solar cells ideally suited to battery charging.

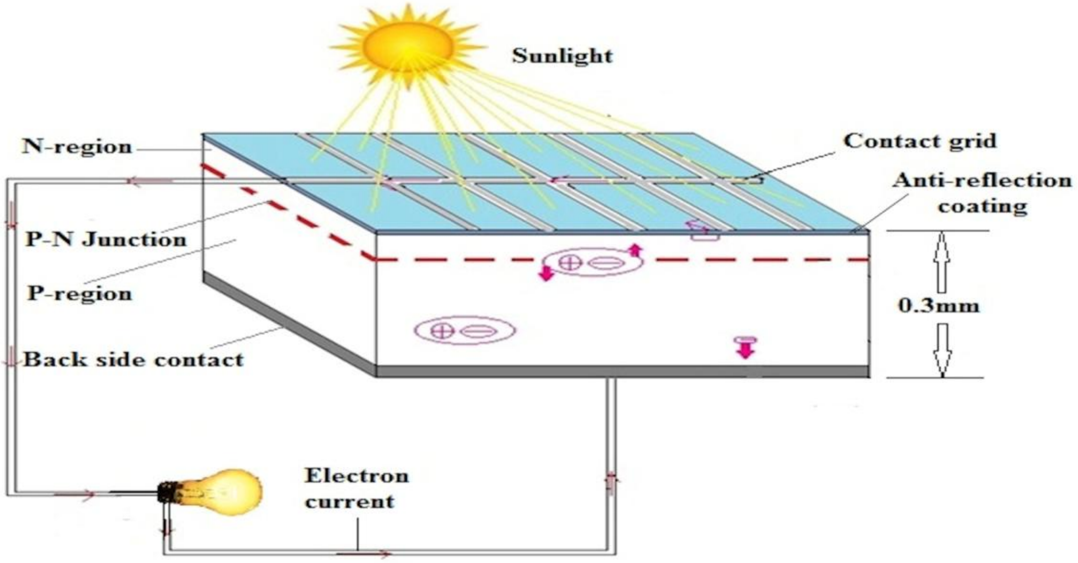


Figure 1: Solar cell device structure

Points to remember:

i) Photon is absorbed and energy is given to an electron in the crystal lattice.

ii) Generated free electrons-hole pair flow through the materials to produce electricity.

iii) Different PV materials have different band gap energies.

iv) Photons with energy equal to the band gap energy are absorbed to create free electrons.

v) Photons with less energy than the band gap energy pass through the material.

2. Electrical characteristics of solar cell

Solar cell is basically a p-n junction diode. It requires no bias across the junction to produce electric current. Surface layer of p type material is made extremely thin to facilitate the incident light to penetrate the junction easily. The device is packaged with glass shutter on the top. When these incident light called photons collide with the valence electrons, they transfer sufficient energy to them to detach the electrons from the parent atoms. Electrons are transferred from the valence band to the conduction band. In this way free electrons and holes are generated on both sides of the junction and their flow constitutes the minatory current IL in the reverse biased direction. In absence of light, thermally generated minority carriers constitute the reverse saturation current. The corresponding I/V characteristic is described by the Shockley solar cell equation

The net p-n junction current which is described by the Shockley solarcell equation,

$$I \ = \mathrm{I_L} - \mathrm{I_F} = \mathrm{I_L} - \mathrm{I_S}[\ \exp\ (\ \mathrm{eV}\ /\ \mathrm{KT})\ -\ 1$$

Where K is the Boltzmann constant, T is the absolute temperature, e(.0) is the electron charge, and V is the voltage at the terminals of thecell. IS is the diode saturation current.

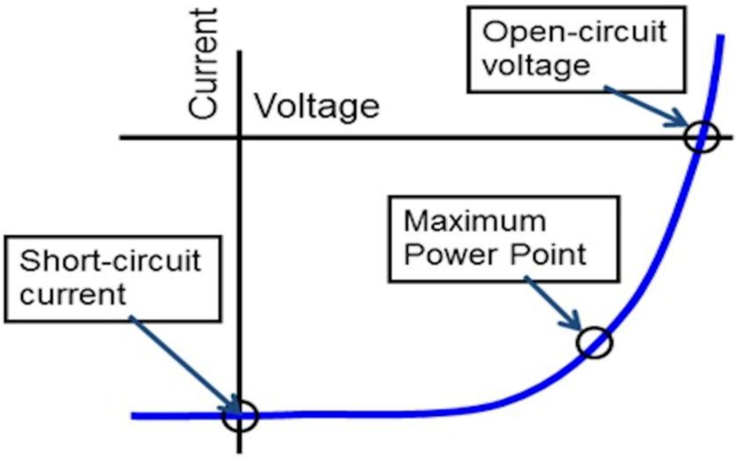


Figure 2: I-V curve

As the diode becomes forward biased due to the voltage drop across the load, therefore magnitude of electric field in space charge region decreases but does not go to zero or change direction. The photocurrent is always in the reverse bias direction. This current is directly proportional to the illumination and also depends on the surface area that is illuminated.

Current –Voltage characteristics:

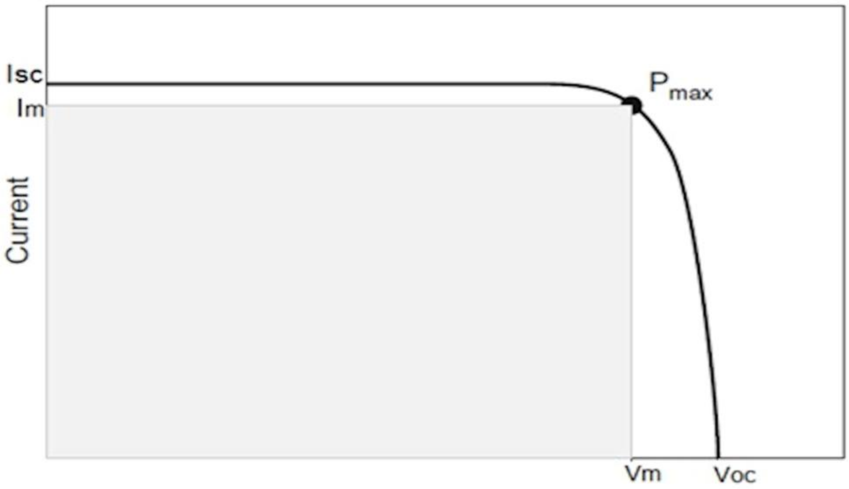


Figure 3: Maximum power rectangle

We observe that Voc is the maximum voltage obtained at the load under open circuit conditions of the diode and Isc is the maximum current through the load under short circuit conditions.

The open circuit voltage Voc is given by:

$$\boldsymbol{Voc} = \frac{\mathbf{KT}}{\mathbf{q}}\mathbf{ln}(\mathbf{1} + \frac{\mathbf{Isc}}{\mathbf{Is}})$$

The power delivered by the device can be maximized by maximizing the area under the curve. This can be done by appropriate selection of load resistor connected across the device.

Both p and n region of the device are heavily doped. Power rectangle curve shows that to have the maximum power, short circuit current and open circuit voltage should be maximum.

There is a term ‘Fill Factor’ (FF) in the theory of solar cell. The Fill Factor is defined as:

FF = Pm/Voc IL

Where, Pm is the maximum output power. Vocis the maximum voltage obtained at the load under open circuit conditions of the diode and Isc is the maximum current through the load under short circuit conditions. Fill Factor depends mainly on Voc/KT. The exact value of Fill Factor can be determined by the following equation:

$$\boldsymbol{FF} = (\mathbf{Voc} - \mathbf{ln}(\mathbf{Voc} + \mathbf{0.72})/(\mathbf{Voc} + \mathbf{1})$$

The above equation shows that the closer the value to unity makes higher the quality of solar cell. The value of FF is generally 0.7 to 0.8.

The conversion efficiency of a solar cell is desired as the ratio of output electrical power to incident optical power.

$$\boldsymbol{\eta} = \frac{\mathbf{Pm}}{\mathbf{Pin}}\ \mathbf{x\ 100\%} = \frac{\mathbf{Im\ Vm}}{\mathbf{Pin}}\ \mathbf{x\ 100\%} = \mathbf{FF}\ \frac{\mathbf{ILVoc}}{\mathbf{Pin}}\ \mathbf{x\ 100\%}$$

The ratio $\frac{\mathrm{Im\ Vm}}{\mathrm{Isc\ Voc}}$ is called Fill Factor and is a measure of reliable power from a solar cell. Where Vm and Im are the voltage and current at the point of maximum power and Pin is the incident optical power.

It is important to note that to realize a high efficiency solar cell, it is not necessary to have high Vocand Isc, the FF value should be also very high. 10-12% efficiency is common for silicon solar cell.

The actual I-V characteristics of solar cell may vary from the ideal one. An equivalent two diode model is used to present the IV curve. The ideality factor of second diode is 2 which is reflected in the second part of the equation (exponential term).

Therefore, the actual one can be written as:

$$\mathbf{I= Isc\text{-}Is1[exp(}\frac{\mathbf{V+IRse}}{\mathbf{KT}}\mathbf{)]} - \mathbf{Is2}\left[\mathbf{exp}\left(\frac{\mathbf{V+IRse}}{\mathbf{2KT}}\right) - \mathbf{1}\right] - \frac{\mathbf{V+IRse}}{\mathbf{Rsh}}$$

In the above figure, it is observed that the circuit contain series resistance (Rse) as well as shunt resistance (Rsh).

How these resistances affect the IV curve of the solar cell is shown in figure.

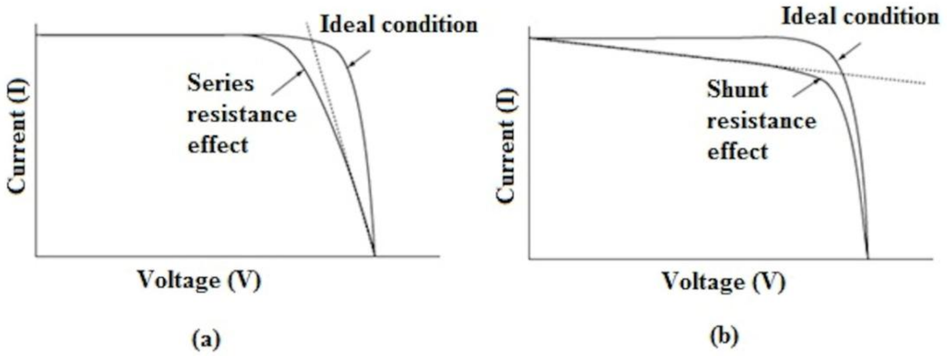


Figure 4: (a) series and (b) parallel resistance effect on I-V characteristic**.**

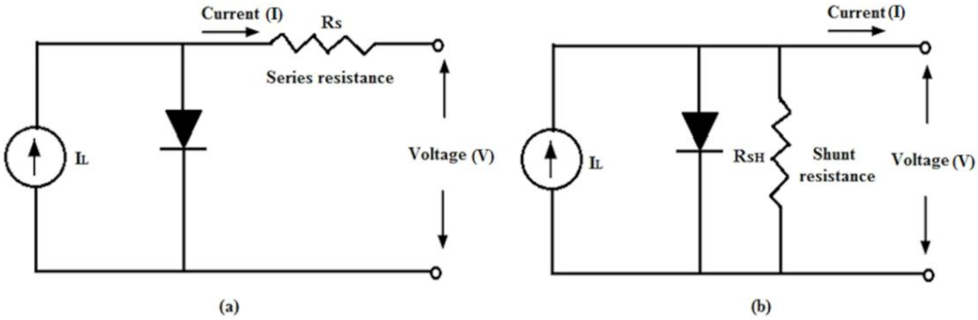


Figure 5: Non ideal solar cell (a) series resistance (b) shunt resistance

The ultimate target of solar cell is to maximize the efficiency by maximizing the absorption and to minimizing the recombination. To maximize the efficiency, the parasitic losses should be minimum. FF and efficiency both are decreased by the shunt and series resistance losses.

Shunt resistance appears in the solar cell due to the processing defect while fabrication , on the other hand series resistance appears due to defects while contact deposition and also their design. Higher shunt resistance is always preferred as low shunt resistance creates an alternating path for the photo current and power loss increases. The exact value of the shunt resistance can be determined from the slope of the I-V curve near the short circuit current point. Typical value shunt resistance is 1000Ωcm^2 for commercial solar cell.

Semiconductor- semiconductor junction: Hetero junction

Energy band diagram, Classification of Hetero Junction, 2D Electron Gas (Isotype Heterojunction), Anisotype Heterojunction, I-V Characteristics. Numerical Problems.

when two different semiconductor materials are used to form a junction, the junction is called a *semiconducror heterojunction.* The hetero junction is formed with narrow bandgap material with wide bandgap material. Figure 1 shows three possible combinations. When the forbidden band gap of wide bandgap material completely overlaps the bandgap of narrow bandgap materials then it is called straddling which is applicable for most of the heterojunctions. In staggered, a fraction of wide bandgap and narrow band gap are merged tog ether. In case of broken gap there is a bandgap itself between energy bandgaps of different semiconductor.

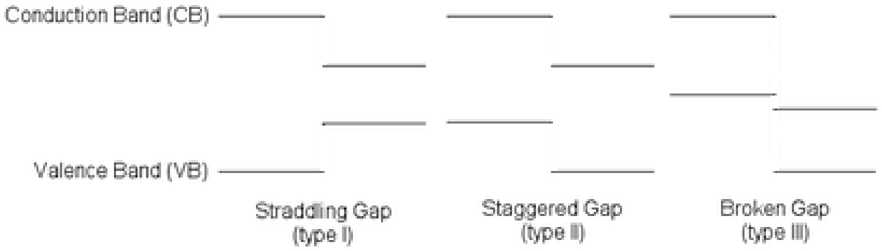
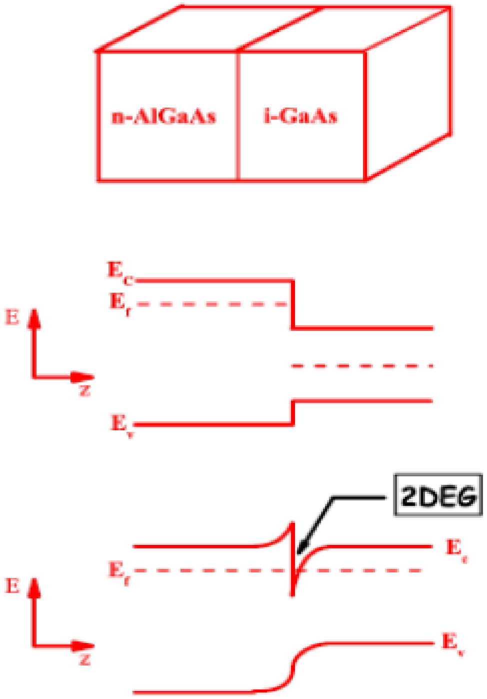
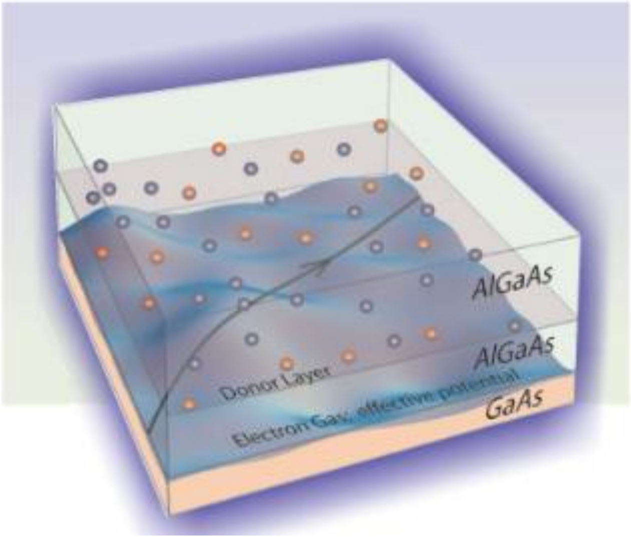


Fig.1. different types of heterojunction (a) straddling (b)staggered (c) broken gap

A two-dimensional electron gas (2DEG) is a scientific model in solid-state physics. It is an electron gas that is free to move in two dimensions, but tightly confined in the third. This tight confinement leads to quantized energy levels for motion in the third direction, which can then be ignored for most problems. Thus the electrons appear to be a 2D sheet embedded in a 3D world. The analogous construct of holes is called a two-dimensional hole gas (2DHG), and such systems have many useful and interesting properties.





Most 2DEG are found in transistor-like structures made from semiconductors. The most commonly encountered 2DEG is the layer of electrons found in MOSFETs. When the transistor is in inversion mode, the electrons underneath the gate oxide are confined to the semiconductor-oxide interface, and thus occupy well defined energy levels. Nearly always, only the lowest level is occupied (see the figure caption), and so the motion of the electrons perpendicular to the interface can be ignored. However, the electron is free to move parallel to the interface, and so is quasi-two-dimensional.

Other methods for engineering 2DEGs are high-electron-mobility-transistors (HEMTs) and rectangular quantum wells. HEMTs are field-effect transistors that utilize the heterojunction between two semiconducting materials to confine electrons to a triangular quantum well. Electrons confined to the heterojunction of HEMTs exhibit higher mobilities than those in MOSFETs, since the former device utilizes an intentionally undoped channel thereby mitigating the deleterious effect of ionized impurity scattering. Two closely spaced heterojunction interfaces may be used to confine electrons to a rectangular quantum well. Careful choice of the materials and alloy compositions allow control of the carrier densities within the 2DEG.

Electrons may also be confined to the surface of a material. For example, free electrons will float on the surface of liquid helium, and are free to move along the surface, but stick to the helium; some of the earliest work in 2DEGs was done using this system.[1] Besides liquid helium, there are also solid insulators (such as topological insulators) that support conductive surface electronic states.

Recently, atomically thin solid materials have been developed (graphene, as well as metal dichalcogenide such as molybdenum disulfide) where the electrons are confined to an extreme degree. The two-dimensional electron system in graphene can be tuned to either a 2DEG or 2DHG by gating or chemical doping. This has been a topic of current research due to the versatile (some existing but mostly envisaged) applications of graphene.[2]

A separate class of heterostructures that can host 2DEGs are oxides. Although both sides of the heterostructure are insulators, the 2DEG at the interface may arise even without doping (which is the usual approach in semiconductors). Typical example is a ZnO/ZnMgO heterostructure.[3] More examples can be found in a recent review[4] including a notable discovery of 2004, a 2DEG at the LaAlO3/SrTiO3 interface[5] which becomes superconducting at low temperatures. The origin of this 2DEG is still unknown, but it may be similar to modulation doping in semiconductors, with electric-field-induced oxygen vacancies acting as the dopants.

2-dimensional electron gas (2DEG) is a scientific model in solid-state physics. It is an electron gas that is free to move in two dimensions, but tightly confined in the third. This tight confinement leads to quantized energy levels for motion in the third direction, which can then be ignored for most problems. Thus the electrons appear to be a 2D sheet embedded in a 3D world. The analogous construct of holes is called a two-dimensional hole gas (2DHG), and such systems have many useful and interesting properties.

Mostly 2DEG are found in transistor-like structures made from semiconductors. The most commonly encountered 2DEG is the layer of electrons found in MOSFETs. When the transistor is in inversion mode, the electrons underneath the gate oxide are confined to the semiconductor-oxide interface, and thus occupy well defined energy levels. Nearly always, only the lowest level is occupied, and so the motion of the electrons perpendicular to the interface can be ignored. However, the electron is free to move parallel to the interface, and so is quasi-two-dimensional.

Other methods for engineering 2DEGs are high-electron-mobility-transistors(HEMTs) and rectangular quantum wells. HEMTs are field-effect transistors that utilize the heterojunction between two semiconducting materials to confine electrons to a triangular quantum well. Electrons confined to the heterojunction of HEMTs exhibit higher mobilities than those in MOSFETs, since the former device utilizes an intentionally undoped channel thereby mitigating the deleterious effect of ionized impurity scattering. Two closely spaced heterojunction interfaces may be used to confine electrons to a rectangular quantum well. Careful choice of the materials and alloy compositions allow control of the carrier densities within the 2DEG.

More examples can be found in a review including a notable discovery of 2004, a 2DEG at the LaAlO3/SrTiO3 interface which becomes superconducting at low temperatures.

Metal semiconductor junction:

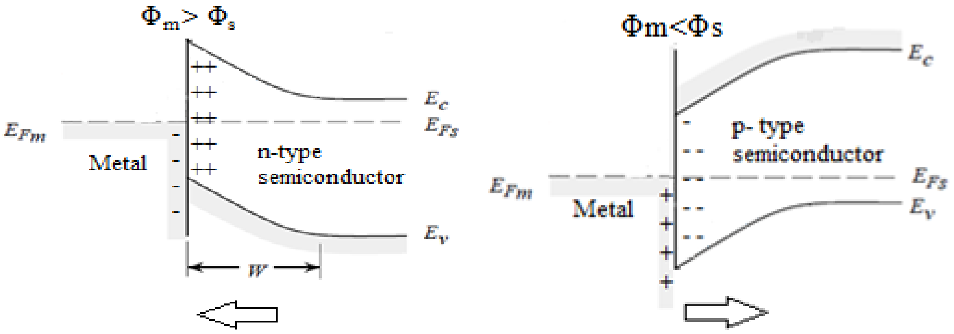
Metal-semiconductor contacts are used in various types of devices. These types of electrical contacts can be of two types. They are Schottky contacts and Ohmic contacts.

In case of the Ohmic contacts, irrespective of the type of biasing condition applied, electrons can flow across the junctions due to the presence of a very small potential barrier at the interface. Whereas in Schottly type of contact, the electrons are capable of flowing across junction under forward bias and their flow is restricted under reverse bias. Thus a Schottky contact acts as a diode and the current flow is determined by the applied voltage.

In order to elaborate on the above mentioned contacts, we must understand some basic terms. In metals, all energy level are filled by electrons upto the Fermi level Ef. The amount of energy qΦm required to remove an electron from the Fermi level to the vacuum is called the work function of the metal. The electron affinity is defined as the energy required to extract an electron from the conduction band of a semiconductor to the vacuum level is called the electron affinity. The band diagrams of the two contact types must be studied in order to get a better understanding of the complete physical phenomena taking place within the materials at the surface of contact.

6.1 Schottky Barriers: A metal with work function qΦm and a semiconductor of work function qΦs is taken such that Φm>Φs. So Efs>Efm. When the metal and the semiconductor is brought in contact, at the equilibrium condition the Fermi levels on both sides of the contact is perfectly aligned. Charge transfer i.e electron flow takes place from the semiconductor to the metal until the Fermi levels are aligned to attain equilibrium. So the electron energy at the side of the semiconductor must be lowered so that the Fermi levels of the two materials are at the same level. A depletion region will be formed consisting of uncompensated donor ions which will be completely balanced by the electrons on the metallic side of the contact. A downward bending of the conduction and the valence band takes place and the contact potential Vo prevents the further electron flow from the semiconductor to the metal.

To this type of barrier, if a forward biasing voltage, V is applied, then the contact potential gets lowered by Vo –V which results in the diffusion of electrons from the conduction band of the semiconductor to metal. This gives rise to a forward current. Conversely, when a reverse biasing voltage Vr is applied, the barrier height gets enhanced by Vo +V and the electron flow is hindered. A very negligible amount of current flows from semiconductor to metal as shown in figure 8. This type of contact is also called rectifying contact.

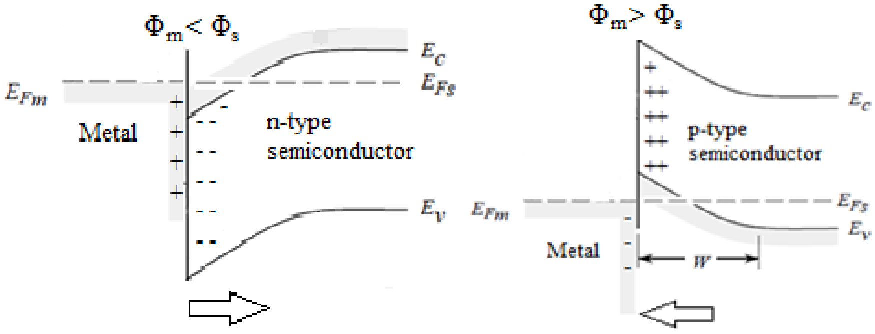


(a) **(b)**

Figure 8. Schottky contacts (a) Φm> Φs (b) Φm< Φs

6.2 Ohmic Barrier: In certain device applications, we may require a metal-semiconductor contact which has a linear I-V characteristic for both types of biasing. Such contacts are called Ohmic contacts. We take a metal and an n-type semiconductor such that Φm<Φs. when these two types of materials are brought in contact with each other then the equilibrium condition is achieved by transfer of electrons from metal to semiconductor. The transfer of charge continues till the Fermi level of both sides of the contact gets aligned to an equal level. Upward bending of the valence and conduction band takes place. Positive space charges are present towards the metallic side of the contact and free electrons are accumulated towards the semiconductor side of the contact.

Under the application of the forward or reverse bias, the barrier height faced by the electrons is small enough to be overcome by them and a considerable amount of current can flow in both directions across the junction and the current follows a linear relationship with the applied voltage (figure 9). The nature of biasing only governs the direction of current flow.



(a) **(b)**

Figure 9. Ohmic contacts (a) Φm> Φs (b) Φm< Φs

In the process of harvesting energy from the NGs, Schottky and Ohmic contacts will be required singly or in combination so that maximum electrical power output can be extracted from the NGs of different designs and materials.

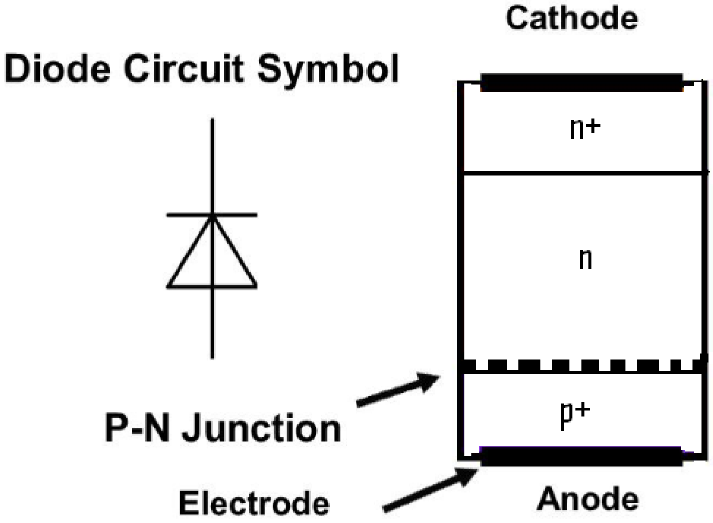
PIN Diode

Although diodes with a basic PN junction are by far the most popular type of diode in use, other types of diode may be used in a number of applications. One type that is used for a variety of circuits is the PIN diode. This form of diode is used in a number of areas. The PIN diode is very good for RF switching, and the PIN structure is also very useful in photodiodes. A further use of the PIN diode is as a photo-detector (photodetector or photo-diode) where its structure is particularly suited to absorbing light.

Structure and Working of PIN Diode

The term PIN diode gets its name from the fact that includes three main layers. Rather than just having a P-type and an N-type layer, it has three layers such as

- P-type layer
- Intrinsic layer
- N-type layer



The working principle of the PIN diode exactly same as a normal diode. The main difference is that the depletion region, because that normally exists between both the P & N regions in a reverse biased or unbiased diode is larger. In any PN junction diode, the P region contains holes as it has been doped to make sure that it has a majority of holes. Likewise the N-region has been doped to hold excess electrons.

PIN diode characteristics

The intrinsic layer between the P-type and N-type regions of the PIN diode enable it to provide properties such as a high reverse breakdown voltage, and a low level of capacitance, and there are also other properties such as carrier storage when it is forward biased that enable it to be used for certain microwave applications.

It is found that at low levels of reverse bias the depletion layer become fully depleted. Once fully depleted the PIN diode capacitance is independent of the level of bias because there is little net charge in the intrinsic layer. However the level of capacitance is typically lower than other forms of diode and this means that any leakage of RF signals across the diode is lower.

When the PIN diode is forward biased both types of current carrier are injected into the intrinsic layer where they combine. It is this process that enables the current to flow across the layer.

The particularly useful aspect of the PIN diode occurs when it is used with high frequency signals, the diode appears as a resistor rather than a non linear device, and it produces no rectification or distortion. Its resistance is governed by the DC bias applied. In this way it is possible to use the device as an effective RF switch or variable resistor producing far less distortion than ordinary PN junction diodes.

PIN diode uses and advantages

The PIN diode is used in a number of areas as a result of its structure proving some properties which are of particular use.

High voltage rectifier: The PIN diode can be used as a high voltage rectifier. The intrinsic region provides a greater separation between the PN and N regions, allowing higher reverse voltages to be tolerated.

RF switch: The PIN diode makes an ideal RF switch. The intrinsic layer between the P and N regions increases the distance between them. This also decreases the capacitance between them, thereby increasing he level of isolation when the diode is reverse biased.

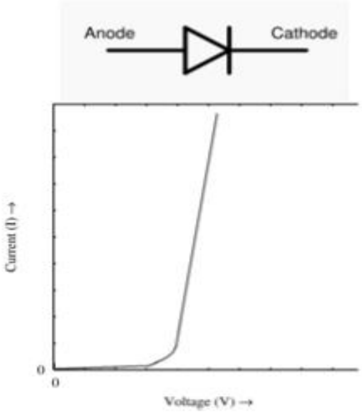
Photodetector: As the conversion of light into current takes place within the depletion region of a photdiode, increasing the depletion region by adding the intrinsic layer improves the performance by increasing he volume in which light conversion occurs.

These are three of the main applications for PIN diodes, although they can also be used in some other areas as well.

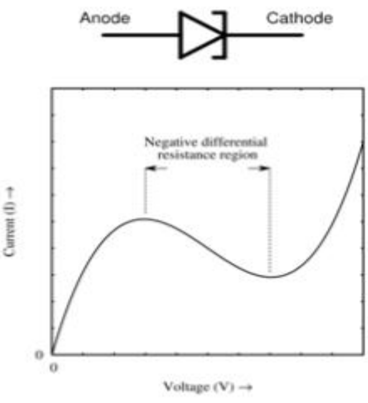
Tunnel Diode (Esaki Diode)

Esaki diodes were named after Leo Esaki, who in 1973 received the Nobel Prize in Physics for discovering the electron tunneling effect used in these diodes,.

Tunnel diode is the p-n junction device that exhibits negative resistance. That means when the voltage is increased the current through it decreases.



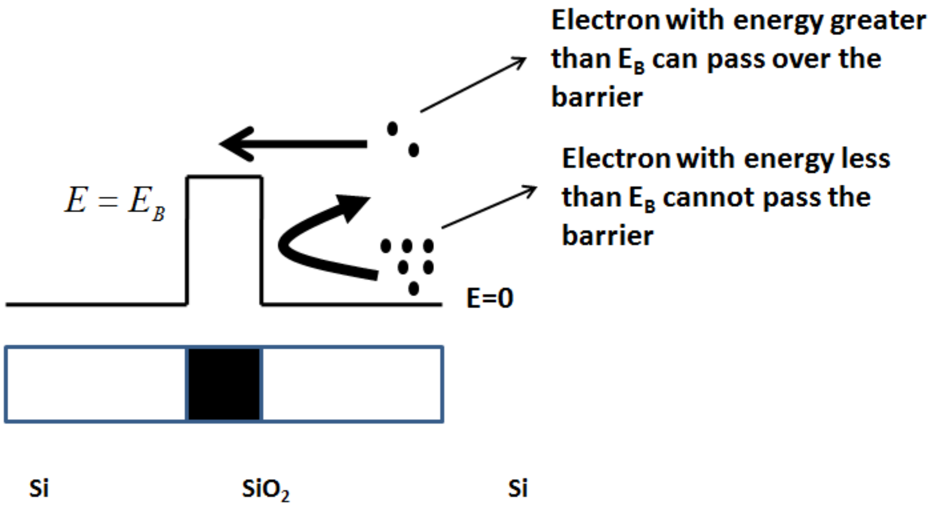
Regular p-n Diode



Tunnel Diode

Concept of Electron Tunneling

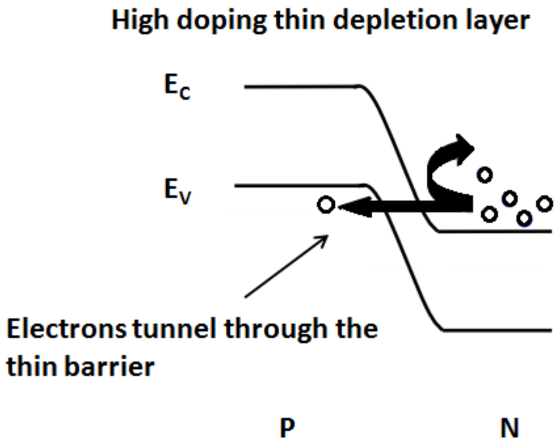
- **For thick barrier, both Newtonian and Quantum mechanics say that the electrons cannot cross the barrier.**
- **It can only pass the barrier if it has more energy than the barrier height.**


Electron with energy greater than E_B can pass over the barrier
Electron with energy less than E_B cannot pass the barrier
$E = E_B$
E=0
Si
SiO_2
Si

- **For thin barrier, Newtonian mechanics still says that the electrons cannot cross the barrier.**
- **However, Quantum mechanics says that the electron wave nature will allow it to tunnel through the barrier.**

Electron Tunneling in p-n junction

When the p and n region are highly doped, the depletion region becomes very thin. In that case, there is a finite probability that electrons can tunnel from the conduction band of n-region to the valence band of p-region. During the tunneling the particle energy does not change.

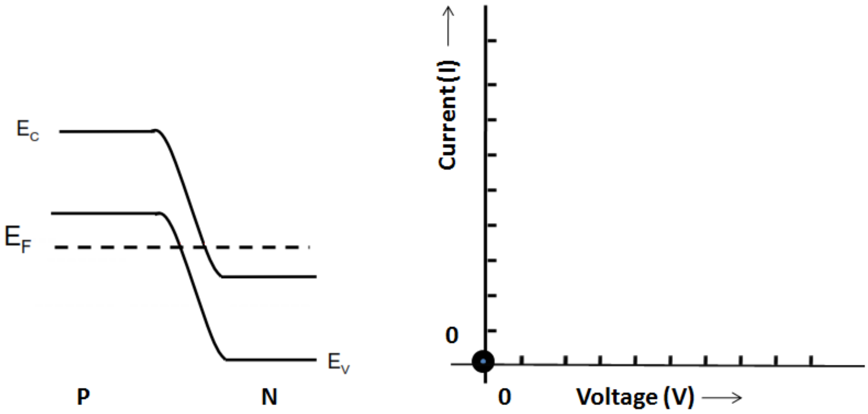

High doping thin depletion layer
E_C
E_V
Electrons tunnel through the thin barrier
P
N

Tunnel Diode Operation:

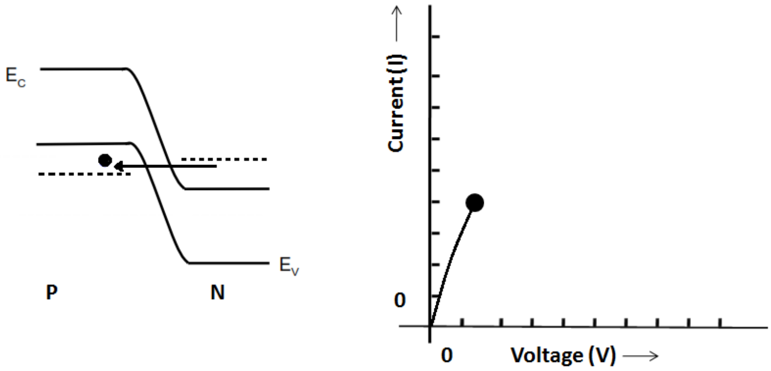
When the semiconductor is very highly doped (i.e. degenerate type of semiconductor) the Fermi level goes above the conduction band for n-type and below valence band for p- type material.
The tunnel diode operates on these degenerate type of materials.

Under Forward Bias:

Step 1: At zero bias there is no current flow

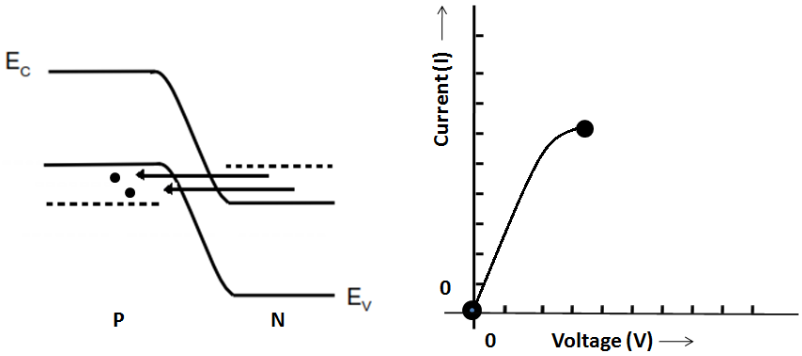


Step 2: A small forward bias is applied. Potential barrier is still very high – no noticeable injection and forward current through the junction.
However, electrons in the conduction band of the n region will tunnel to the empty states of the valence band in p region. This will create a forward bias tunnel current.



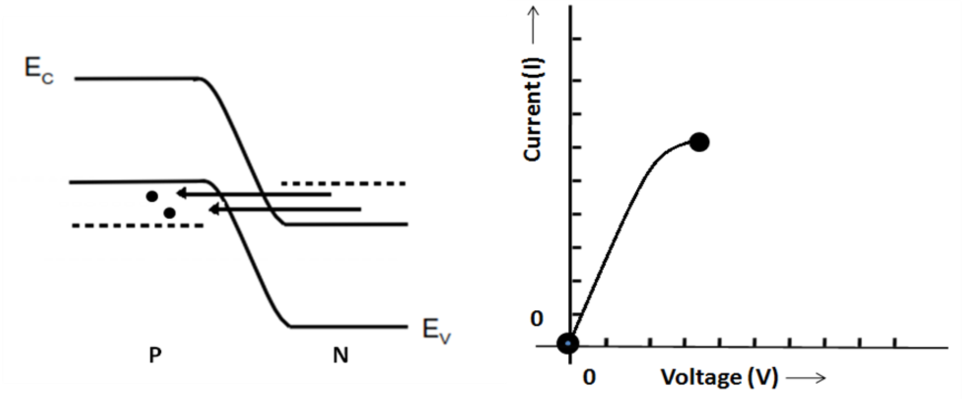
Direct tunneling current starts growing

Step 3: With a larger voltage the energy of the majority of electrons in the n-region is equal to that of the empty states (holes) in the valence band of p-region; this will produce maximum tunneling current



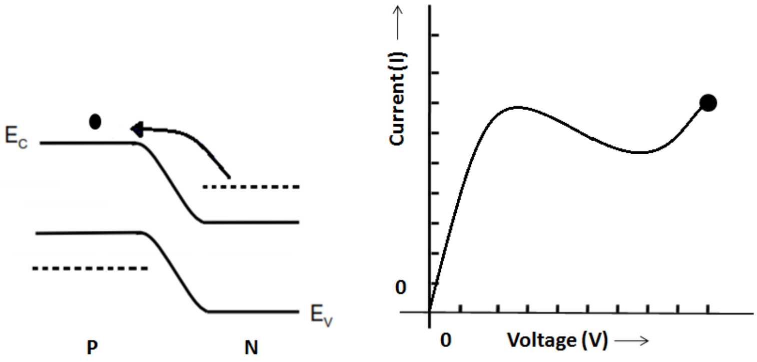
Maximum Direct tunneling current

Step 4: As the forward bias continues to increase, the number of electrons in the n side that are directly opposite to the empty states in the valence band (in terms of their energy) decrease. Therefore decrease in the tunneling current will start.



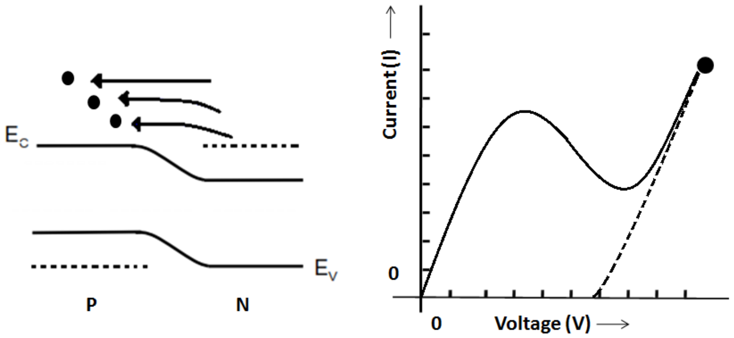
Direct tunneling current decreases

Step 5: As more forward voltage is applied, the tunneling current drops to zero. But the regular diode forward current due to electron – hole injection increases due to lower potential barrier.



No tunneling current; diffusion current starts growing

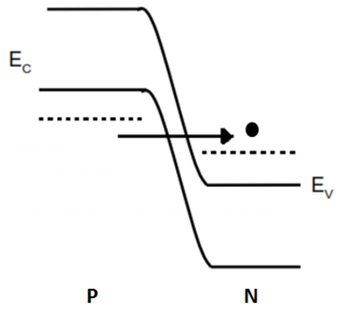
Step 6: With further voltage increase, the tunnel diode I-V characteristic is similar to that of a regular p-n diode.

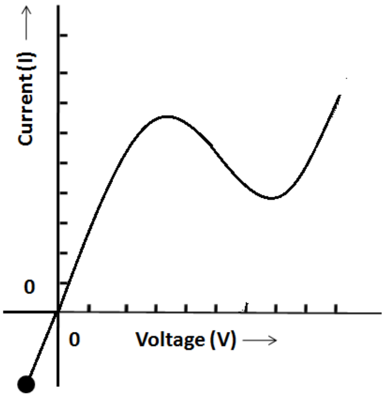


Under Reverse Bias

In this case the, electrons in the valence band of the p side tunnel directly towards the empty states present in the conduction band of the n side creating large tunneling current which increases with the application of reverse voltage.

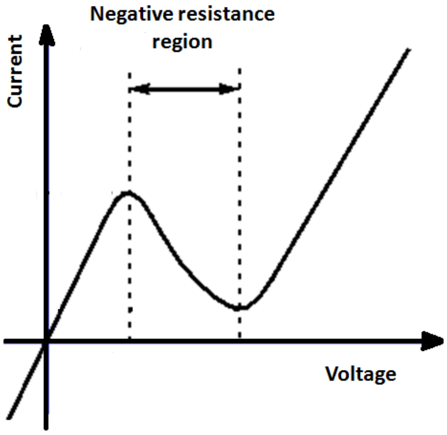
The Tunnel Diode reverse I-V is similar to the Zener diode with nearly zero breakdown voltage.





Negative resistance of Tunnel diode

The characteristic curve for a tunnel diode shows an area of negative resistance. When forward biased the current in the diode rises at first, but later it can be seen to fall with increasing voltage, before finally rising again.



SILICON CONTROLLED RECTIFIERS (SCR)

A silicon controlled rectifier is a semiconductor device that acts as a controlled switch to perform various functions such as rectification, inversion and regulation of power flow. It can change alternating current and at the same time can control the amount of power fed to the load. SCR combines the features of a rectifier and a transistor.

Like the diode, SCR is a unidirectional device, i.e. it will only conduct current in one direction only, but unlike a diode, the SCR can be made to operate as either an open-circuit switch or as a rectifying diode depending upon how its gate is triggered.

In other words, SCR can operate only in the switching mode and cannot be used for amplification.

Hence, it is extensively used in switching d.c. and a.c., rectifying a.c. to give controlled output, converting d.c. into a.c. etc.

CONSTRUCTION

When a an ordinary rectifier (pn) and a junction transistor (npn) combined in one unit to form pnpn device , the resulting three pn junctions device is called a silicon controlled rectifier.

Fig.1(a) shows the construction of an SCR and Fig.1(b) shows the symbol of SCR

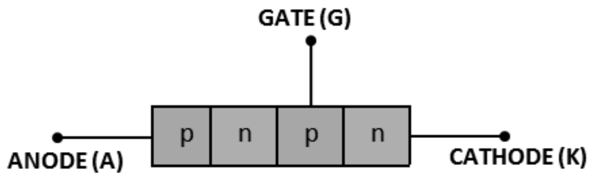


Fig.1(a)

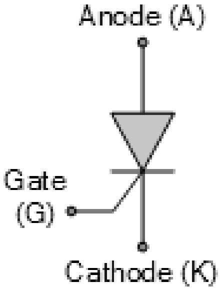


Fig.1(b)

Three terminals are taken; one from the outer p-type material called anode A, second from the outer layer of n-type material called cathode K and the third from the base of transistor section and is called gate G.

In the normal operating conditions of SCR, anode is held at high positive potential w.r.t. cathode and gate at small positive potential w.r.t. cathode.

Working principle of SCR

In a silicon controlled rectifier, load is connected in series with anode. The anode is always kept at positive potential w.r.t. cathode.

The working principle of SCR is processed under two sections:

1. **When gate is open:**

Fig.2 shows the SCR circuit with gate open i.e. no voltage applied to the gate.

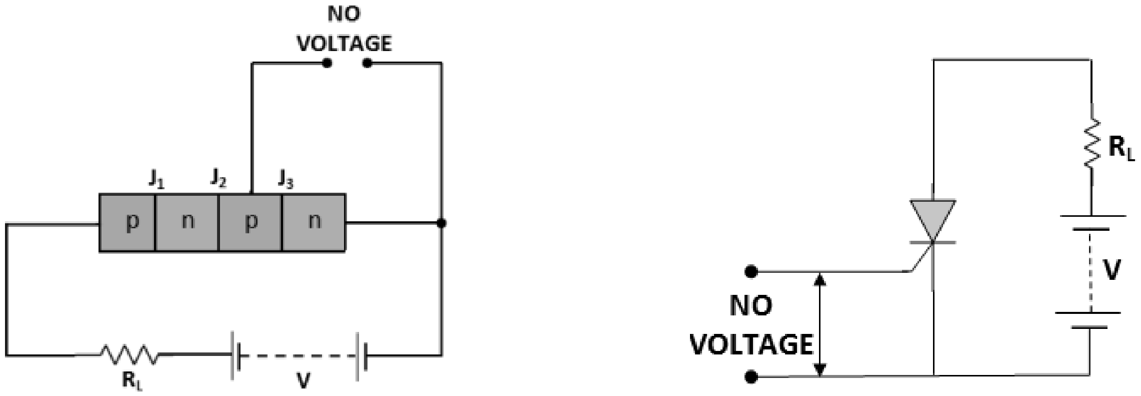


Fig.2

Under this condition, junction J_2 is reverse biased while junction J_1 and J_3 are forward biased.

Hence, the situation in the junctions J_1 and J_3 is just as in a npn transistor with base open.

Consequently, no current flows through the load R_L and the SCR is cut off.

However, if the applied voltage is gradually increased, a stage is reached when the reverse biased junction J_2 breaks down.

The SCR now conducts heavily and is said to be in the ON state.

The applied voltage at which SCR conducts heavily without gate voltage is called Breakover voltage.

2. When gate is positive w.r.t. cathode

The SCR can be made to conduct heavily at smaller applied voltage by applying a small positive potential to the gate as shown in fig.3.

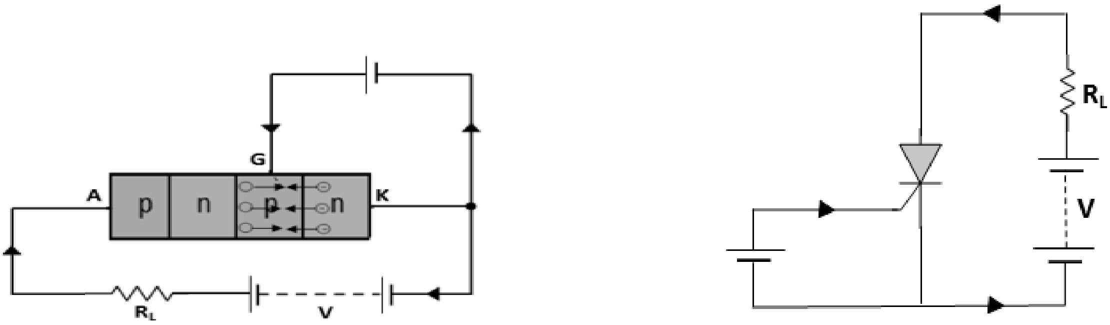


Fig.3

Now junction J_3 is forward biased and junction J_2 is reverse biased.

The electrons from n-type material start moving across junction J_3 towards left whereas holes from p-type towards the right.

Consequently, the electrons from junction J_3 are attracted across the junction J_2 and gate current starts flowing. As soon as the gate current flows, anode current increases.

The increased current in turn makes more electrons available at junction J_2.

This process continues and in an extremely small time, junction J_2 breaks down and the SCR starts conducting heavily.

Once SCR starts conducting, the gate loses all control. Even if gate voltage is removed, the anode current does not decrease at all.

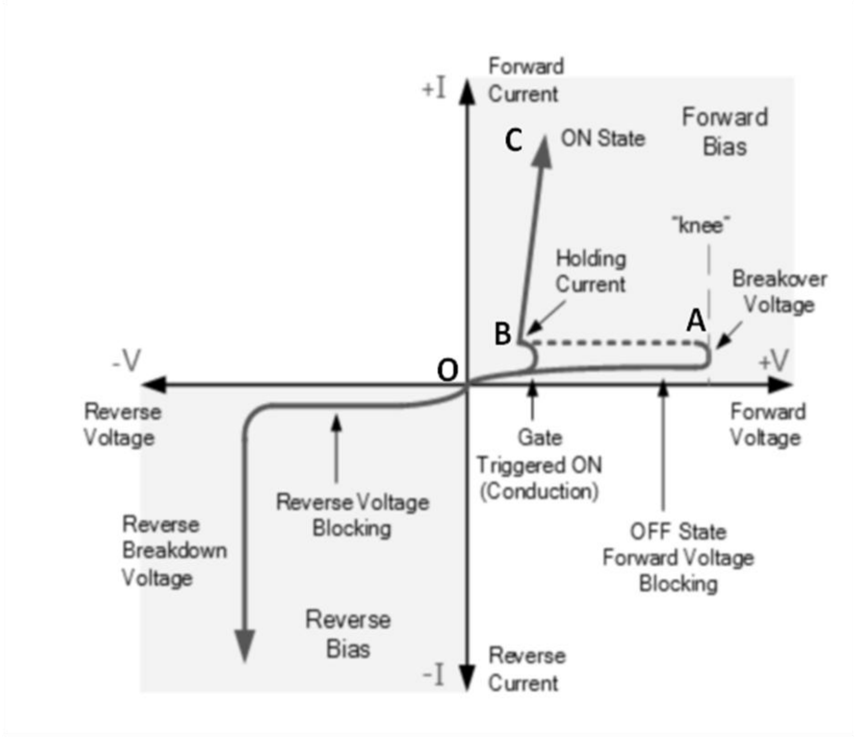
The only way to stop conduction i.e. to bring the SCR in off condition, is to reduce the applied voltage to zero.

V-I Characteristics of SCR

Here is the V-I characteristics of SCR:

It is the curve between anode-cathode voltage (V) and anode current (I) of an SCR at constant gate current.

Fig, 4 shows the V-I characteristics of a typical SCR



The V-I Characteristics of SCR

Forward Characteristics

When anode is positive w.r.t. cathode, the curve between V and I is called the forward characteristics.

In fig.4, OABC is the forward characteristics of SCR at I_G=0.

If the supply voltage is increased from zero, a point reached (point A) when the SCR starts conducting.

Under this condition, the voltage across SCR suddenly drops as shown by dotted curve AB and most of supply voltage appears across the load resistance RL .

If proper gate current is made to flow, SCR can close at much smaller supply voltage.

Reverse Characteristics

When anode is negative w.r.t. cathode, the curve between V and I is known as reverse characteristics.

The reverse voltage does come across SCR when it is operated with a.c. supply.

If the reverse voltage is gradually increased, at first the anode current remains small (i.e. leakage current) and at some reverse voltage, avalanche breakdown occurs and the SCR starts conducting heavily in the reverse direction as shown by the curve DE.

This maximum reverse voltage at which SCR starts conducting heavily is known as reverse breakdown voltage.

Important Terms in The V-I Characteristics of SCR

The following terms are much used in the study of SCR :

1. Breakover voltage
2. Peak reverse voltage
3. Holding current
4. Forward current rating

1. Breakover Voltage

It is the minimum forward voltage, gate being open, at which SCR starts conducting heavily i.e. turned on.

Thus, if the breakover voltage of an SCR is 200 V, it means that it can block a forward voltage (i.e. SCR remains open) as long as the supply voltage is less than 200 V. If the supply voltage is more than this value, then SCR will be turned on.

In practice, the SCR is operated with supply voltage less than breakover voltage and it is then turned on by means of a small voltage applied to the gate.

2. Peak Reverse Voltage (PRV)

It is the maximum reverse voltage (cathode positive w.r.t. anode) that can be applied to an SCR without conducting in the reverse direction.

PRV is an important consideration while connecting an SCR in an a.c. circuit. During the negative half of a.c. supply, reverse voltage is applied across SCR. If PRV is exceeded, there may be avalanche breakdown and the SCR will be damaged if the external circuit does not limit the current.

3. Holding Current

It is the maximum anode current, gate being open, at which SCR is turned OFF from ON condition.

When SCR is in the conducting state, it cannot be turned OFF even if gate voltage is removed.

The only way to turn off or open the SCR is to reduce the supply voltage to almost zero at which point the internal transistor comes out of saturation and opens the SCR.

The anode current under this condition is very small (a few mA) and is called holding current.

Thus, if an SCR has a holding current of 5mA, it means that if anode current is made less than 5 mA, then SCR will be turned off.

4. Forward Current Rating

It is the maximum anode current that an SCR is capable of passing without destruction.

Every SCR has a safe value of forward current which it can conduct. If the value of current exceeds this value, the SCR may be destroyed due to intensive heating at the junction.

For example, if an SCR has a forward current rating of 40 A, it means that the SCR can safely carry only 40 A. Any attempt to exceed this value will result in the destruction of the SCR.

The Triac

A triac is a three-terminal semiconductor switching device which can control alternating current in a load. Triac is an abbreviation for triode a.c. switch. ‘Tri’– indicates that the device has three terminals and ‘ac’ means that the device controls alternating current or can conduct current in either direction.

The key function of a triac may be understood by referring to the simplified Fig. 1(a). The control circuit of triac can be adjusted to pass the desired portions of positive and negative halfcycle of a.c. supply through the load RL. Thus referring to Fig. 1 (b), the triac passes the positive half-cycle of the supply from θ1 to 180° i.e. the shaded portion of positive half-cycle. Similarly, the shaded portion of negative half-cycle will pass through the load. In this way, the alternating current and hence a.c. power flowing through the load can be controlled.

Since a triac can control conduction of both positive and negative half-cycles of a.c. supply, it is sometimes called a bidirectional semi-conductor triode switch.

○○○

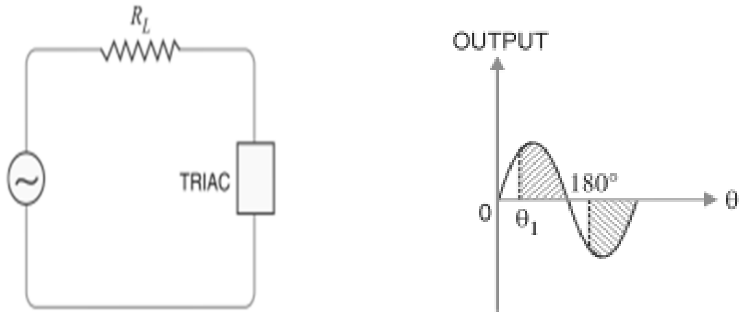


Fig 1 (a) **(b)**

Triac Construction:

A triac is a three-terminal, five-layer semiconductor device whose forward and reverse characteristics are identical to the forward characteristics of the SCR.

The three terminals are designated as main terminal T1, main terminal T2 and gate G. Fig. 2.(a) shows the basic structure of a triac. As we shall see, a triac is equivalent to two separate SCRs connected in inverse parallel (i.e. anode of each connected to the cathode of the other) with gates commoned as shown in Fig. 2(b). Therefore, a triac acts like a bidirectional switch i.e. it can conduct current in either direction. This is unlike an SCR which can conduct current only in one direction. Fig. 2(c) shows the schematic symbol of a triac. The symbol consists of two parallel diodes connected in opposite directions with a single gate lead. It can be seen that even the symbol of triac indicates that it can conduct current for either polarity of the main terminals (T1 and T2) i.e. it can act as a bidirectional switch. The gate provides control over conduction in either direction.

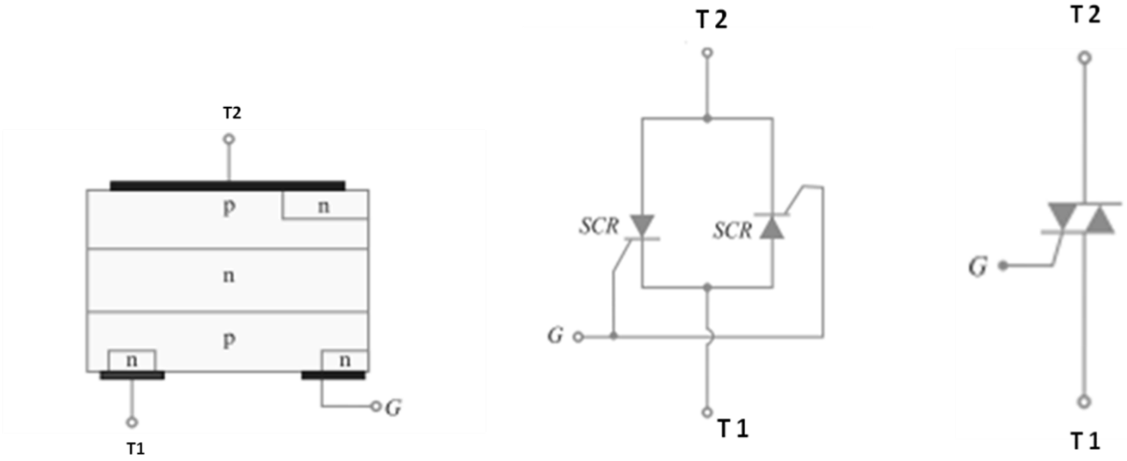


Fig 2(a) (b) (c)

Traic Characteristics:

Fig.3 shows the V-I characteristics of a triac. Because the triac essentially consists of two SCRs of opposite orientation fabricated in the same crystal, its operating characteristics in the first and third quadrants are the same except for the direction of applied voltage and current flow. The following points may be noted from the triac characteristics:

(i) The V-I characteristics for triac in the I st and III rd quadrants are essentially identical to those of an SCR in the I st quadrant.

(ii) The triac can be operated with either positive or negative gate control voltage but in normal operation usually the gate voltage is positive in quadrant I and negative in quadrant III.

(iii) The supply voltage at which the triac is turned ON depends upon the gate current. The greater the gate current, the smaller the supply voltage at which the triac is turned on. This permits to use a triac to control a.c. power in a load from zero to full power in a smooth and continuous manner with no loss in the controlling device.

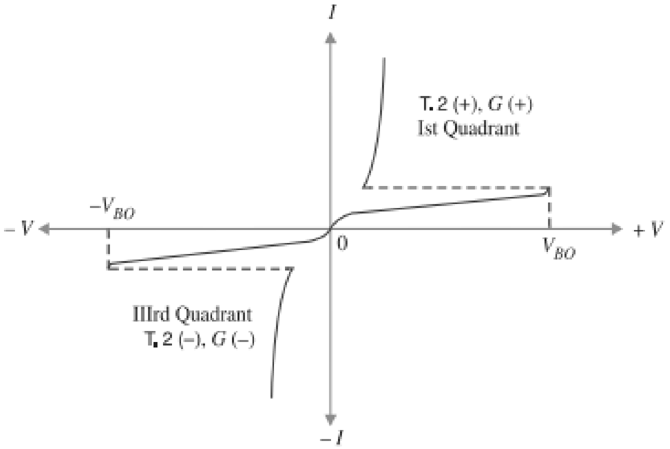
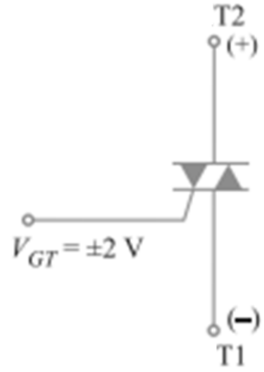


Fig 3

Example 1. The triac shown in Fig. 21.14 can be triggered by the gate triggering voltage VGT = ± 2V. How will you trigger the triac by (i) only a positive gate voltage (ii) only a negative gate voltage?

Solution:



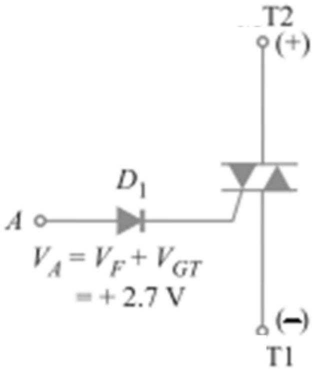


Fig 1 **Fig 2**

In Fig.1. the triac will be triggered into conduction for $V_{GT} = \pm 2V$.

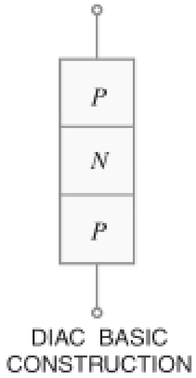
(i) In order that the triac is triggered only by a positive gate voltage, we can use the method shown in Fig. 2. In this circuit, diode D_1 is forward biased when V_{GT} is positive and reverse biased when V_{GT} is negative. Since D_1 will conduct only when V_{GT} is positive, the triac can only be triggered by a positive gate signal. The voltage V_A required to trigger the device is equal to the sum of V_F for diode D_1 and the required gate triggering voltage i.e.

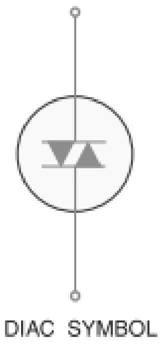
$V_A = V_F + V_{GT} = 0.7V + 2V = 2.7V$

(ii) In order that the triac is triggered only by the negative voltage, reverse the direction of diode D1.

The Diac

A diac is a two-terminal, three layer bidirectional device which can be switched from its OFF state to ON state for either polarity of applied voltage.





(a) **(b)**

Fig 1

The diac can be constructed in either NPN or PNP form. Fig. 1(a) shows the basic structure of a diac in PNP form. The two leads are connected to p-regions of silicon separated by an n-region. The structure of diac is very much similar to that of a transistor. However, there are several important differences:

(i) There is no terminal attached to the base layer.

(ii) The three regions are nearly identical in size.

(iii) The doping concentrations are identical (unlike a bipolar transistor) to give the device symmetrical properties.

Fig. 1 (b) shows the symbol of a diac.

Operation: When a positive or negative voltage is applied across the terminals of a diac, only a small leakage current I_{BO} will flow through the device. As the applied voltage is increased, the leakage current will continue to flow until the voltage reaches the breakover voltage $V_{BO.}$ At this point, avalanche breakdown of the reverse-biased junction occurs and the device exhibits negative resistance i.e. current through the device increases with the decreasing values of applied voltage. The voltage across the device then drops to ‘breakback’ voltage V_W.

Fig. 2 shows the V-I characteristics of a diac. For applied positive voltage less than + V_{BO} and negative voltage less than – V_{BO}, a small leakage current (± I_{BO}) flows through the device. Under such conditions, the diac blocks the flow of current and effectively behaves as an open circuit. The voltages + V_{BO} and – V_{BO} are the breakdown voltages and usually have a range of 30 to 50 volts.

When the positive or negative applied voltage is equal to or greater than the breakdown voltage, diac begins to conduct and the voltage drop across it becomes a few volts. Conduction then continues until the device current drops below its holding current. Note that the breakover voltage and holding current values are identical for the forward and reverse regions of operation.

Diacs are used primarily for triggering of triacs in adjustable phase control of a.c. mains power. Some of the circuit applications of diac are (i) light dimming (ii) heat control and (iii) universal motor speed control.

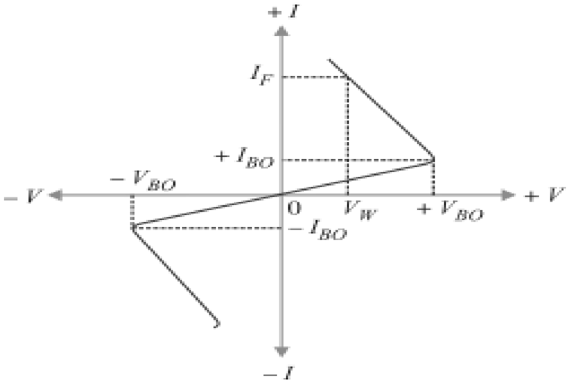


Fig 2

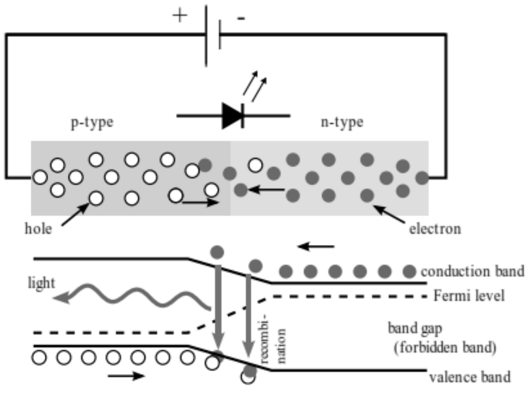
LIGHT EMITTING DIODE

Light Emitting Diode:

The pn junction diode, which is specially doped and made of special type of semiconductor, emits light when it is forward biased is called light emitting diode.

A P-N junction can convert absorbed light energy into a proportional electric current. The same process is reversed here (i.e. the P-N junction emits light when electrical energy is applied to it). This phenomenon is generally called electroluminescence, which can be defined as the emission of light from a semiconductor under the influence of an electric field. The charge carriers recombine in a forward-biased P-N junction as the electrons cross from the N-region and recombine with the holes existing in the P-region. Free electrons are in the conduction band of energy levels, while holes are in the valence energy band. Thus the energy level of the holes is less than the energy levels of the electrons. Some portion of the energy must be dissipated to recombine the electrons and the holes. This energy is emitted in the form of heat and light.

The electrons dissipate energy in the form of heat for silicon and germanium diodes but in gallium arsenide phosphide (GaAsP) and gallium phosphide (GaP) semiconductors, the electrons dissipate energy by emitting photons. If the semiconductor is translucent, the junction becomes the source of light as it is emitted, thus becoming a light-emitting diode. However, when the junction is reverse biased, the LED produces no light and—if the potential is great enough, the device is damaged.

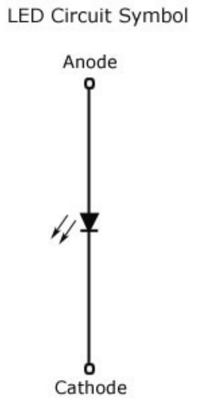


Applications of Light Emitting Diodes

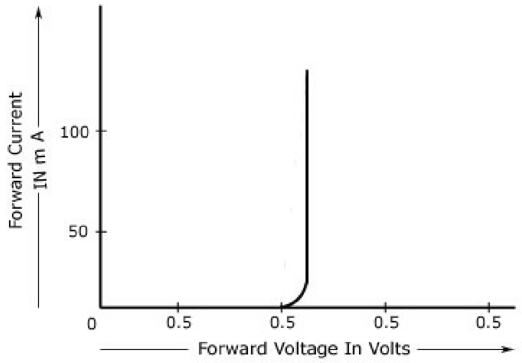
There are many applications of the LED and some of them are explained below.

- LED is used as a bulb in the homes and industries
- The light emitting diodes are used in the motorcycles and cars
- These are used in the mobile phones to display the message
- At the traffic light signals led’s are used

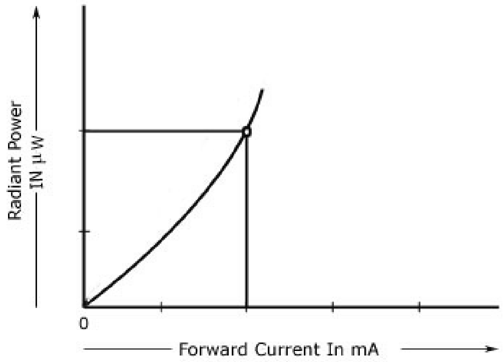
LED Circuit Symbol:



LED Characteristics:



Forward Bias V-I Curve



Output characteristics Curve

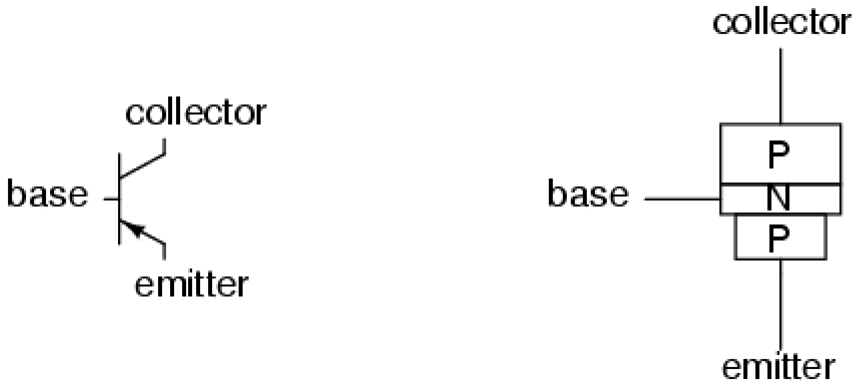
Module IV:BIPOLAR JUNCTION TRANSISTOR

Introduction

The bipolar junction transistor (BJT) was the first solid-state amplifier element and started the solid-state electronics revolution. Bardeen, Brattain and Shockley, while at Bell Laboratories, invented it in 1948 as part of a post-war effort to replace vacuum tubes with solid-state devices. Solid-state rectifiers were already in use at the time and were preferred over vacuum diodes because of their smaller size, lower weight and higher reliability. A solid-state replacement for a vacuum triode was expected to yield similar advantages. The work at Bell Laboratories was highly successful and

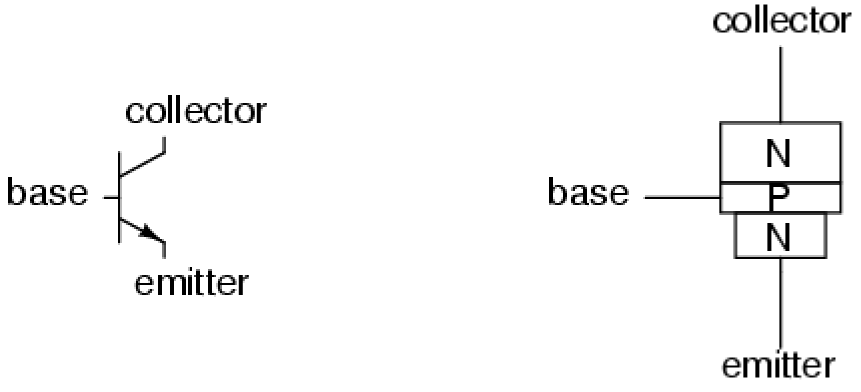
A bipolar transistor consists of a three-layer "sandwich" of doped (extrinsic) semiconductor materials, either P-N-P or N-P-N. Each layer forming the transistor has a specific name, and each layer is provided with a wire contact for connection to a circuit. Shown here are schematic symbols and physical diagrams of these two transistor types:

***PNP* transistor**



schematic symbol *physical diagram*

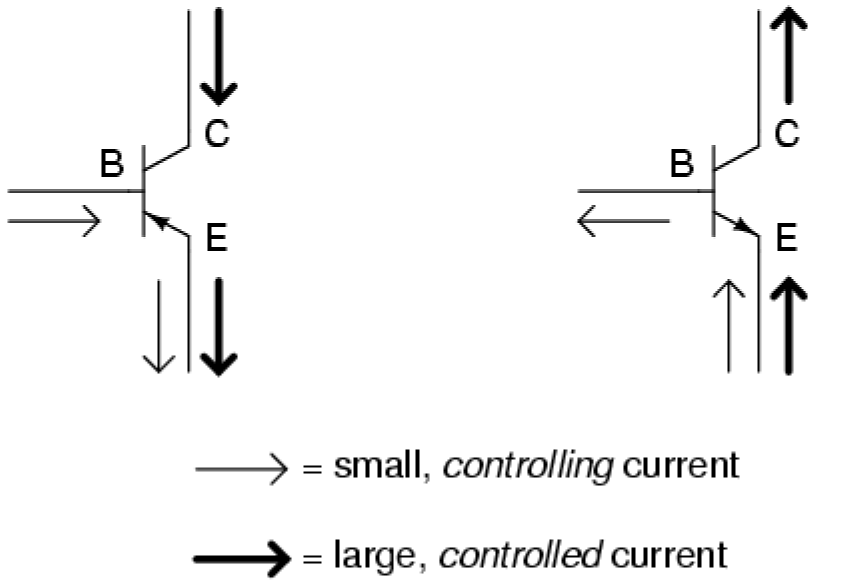
***NPN* transistor**



schematic symbol *physical diagram*

The only functional difference between a PNP transistor and an NPN transistor is the proper biasing (polarity) of the junctions when operating. For any given state of operation, the current directions and voltage polarities for each type of transistor are exactly opposite each other.

Bipolar transistors work as current-controlled current *regulators*. In other words, they restrict the amount of current that can go through them according to a smaller, controlling current. The main current that is *controlled* goes from collector to emitter, or from emitter to collector, depending on the type of transistor it is (PNP or NPN, respectively). The small current that *controls* the main current goes from base to emitter, or from emitter to base, once again depending on the type of transistor it is (PNP or NPN, respectively). According to the confusing standards of semiconductor symbology, the arrow always points *against* the direction of electron flow:



Bipolar transistors are called *bi*polar because the main flow of electrons through them takes place in *two* types of semiconductor material: P and N, as the main current goes from emitter to collector (or visa-versa). In other words, two types of charge carriers -- electrons and holes -- comprise this main current through the transistor.

As you can see, the *controlling* current and the *controlled* current always mesh together through the emitter wire, and their electrons always flow *against* the direction of the transistor's arrow. This is the first and foremost rule in the use of transistors: all currents must be going in the proper directions for the device to work as a current regulator. The small, controlling current is usually referred to simply as the *base current* because it is the only current that goes through the base wire of the transistor. Conversely, the large, controlled current is referred to as the *collector current* because it is the only current that goes through the collector wire. The emitter current is the sum of the base and collector currents, in compliance with Kirchhoff's Current Law.

If there is no current through the base of the transistor, it shuts off like an open switch and prevents current through the collector. If there is a base current, then the transistor turns on like a closed switch and allows a proportional amount of current through the collector. Collector current is primarily limited by the base current, regardless of the amount of voltage available to push it. The next section will explore in more detail the use of bipolar transistors as switching elements.

TRANSISTOR THEORY

You should recall from an earlier discussion that a forward-biased

PN junction is comparable to a low-resistance circuit element because it passes a high current for a given voltage. In turn, a reverse-biased PN junction is comparable to a high-resistance circuit element. By using the Ohm's law formula for power ($P = I^2R$) and assuming current is held constant, you can conclude that the power developed across a high resistance is greater than that developed across a low resistance. Thus, if a crystal were to contain two PN junctions (one forward-biased and the other reverse-biased), a low-power signal could be injected into the forward-biased junction and produce a high-power signal at the reverse-biased junction. In this manner, a power gain would be obtained across the crystal. This concept, which is merely an extension of the material covered in chapter 1, is the basic theory behind how the transistor amplifies. With this information fresh in your mind, let's proceed directly to the NPN transistor.

NPN Transistor Operation:

Just as in the case of the PN junction diode, the N material comprising the two end sections of the NP N transistor contains a number of free electrons, while the center P section contains an excess number of holes. The action at each junction between these sections is the same as that previously described for the diode; that is, depletion regions develop and the junction barrier appears. To use the transistor as an amplifier, each of these junctions must be modified by some external bias voltage. For the transistor to function in this capacity, the first PN junction (emitter-base junction) is biased in the forward, or low-resistance, direction. At the same time the second PN junction (base-collector junction) is biased in the reverse, or high-resistance, direction. A simple way to remember how to properly bias a transistor is to observe the NPN or PNP elements that make up the transistor. The letters of these elements indicate what polarity voltage to use for correct bias. For instance, notice the NPN transistor below:

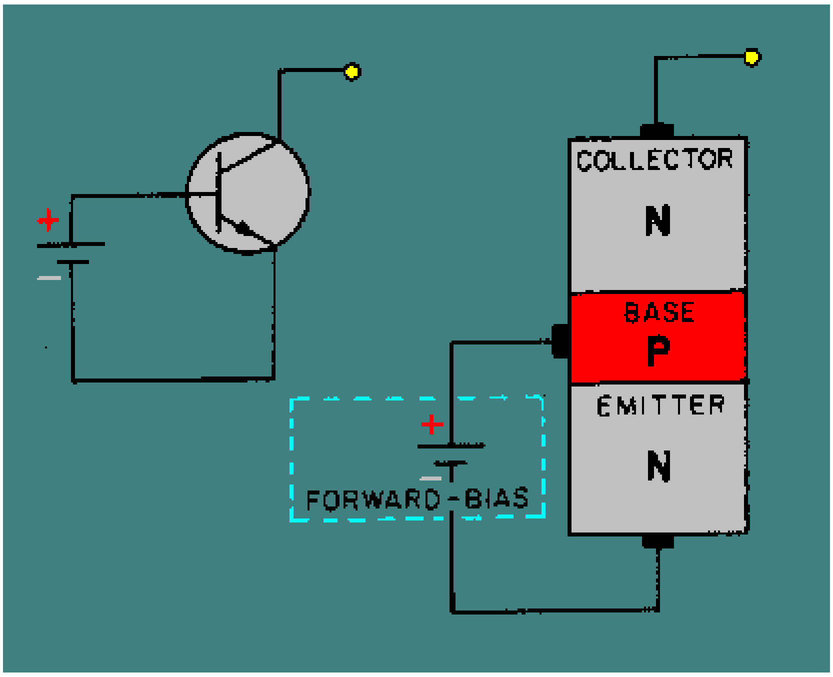


Fig -5.1.1

The emitter, which is the first letter in the NPN sequence, is connected to the negative side of the battery while the base, which is the second letter(NPN), is connected to the positive side. However, since the second PN junction is required to be reverse biased for proper transistor operation, the collector must be connected to an opposite polarity voltage(positive) than that indicated by its letter designation(NPN). The voltage on the collector must also be more positive than the base, as shown below:

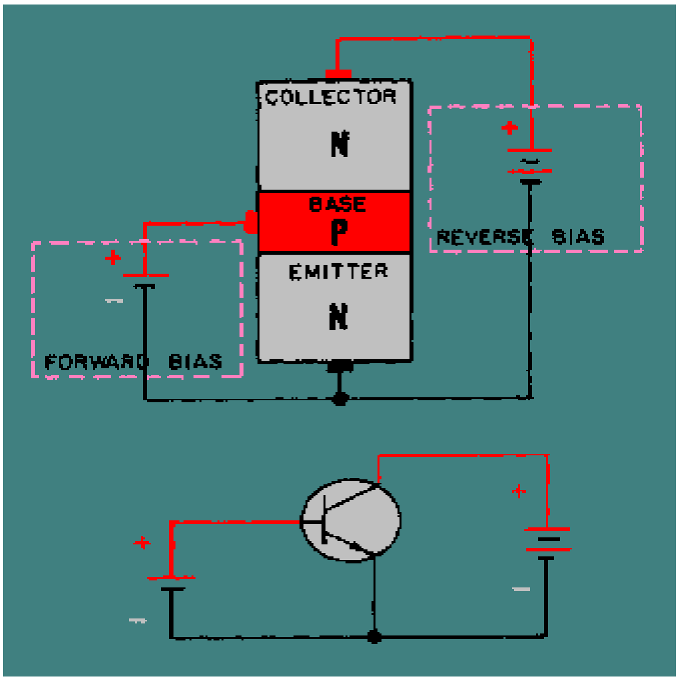


Fig-5.1.2

We now have a properly biased NPN transistor.

In summary, the base of the NPN transistor must be positive with respect to the emitter, and the collector must be more positive than the base.

NPN FORWARD-BIASED JUNCTION. - An important point to bring out at this time, which was not necessarily mentioned during the explanation of the diode, is the fact that the N material on one side of the forward-biased junction is more heavily doped than the P material. This results in more current being carried across the junction by the majority carrier electrons from the N material than the majority carrier holes from the P material. Therefore, conduction through the forward-biased junction, as shown in figure 6.1.3, is mainly by majority carrier electrons from the N material (emitter).

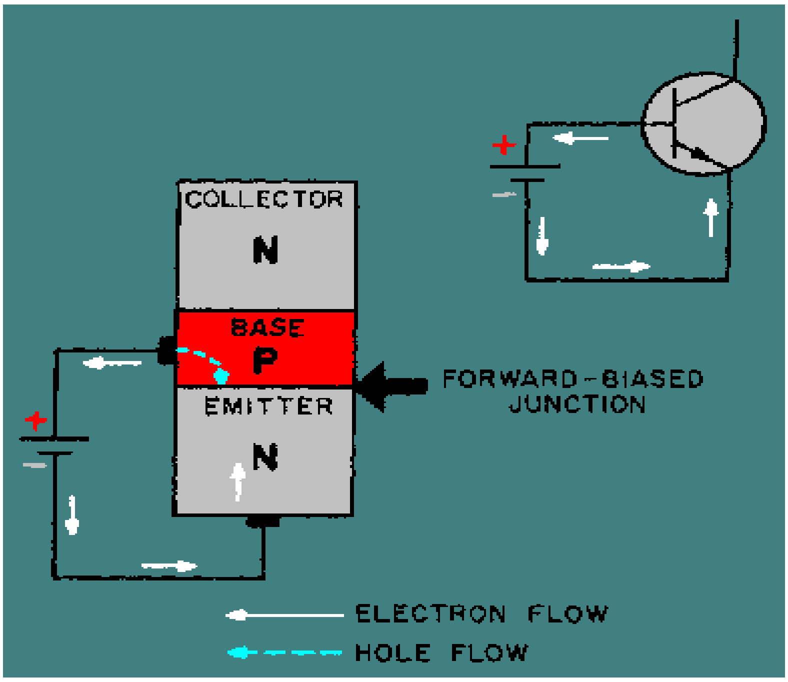


Figure 5.1.3. - The forward-biased junction in an NPN transistor.

With the emitter-to-base junction in the figure biased in the forward direction, electrons leave the negative terminal of the battery and enter the N material (emitter). Since electrons are majority current carriers in the N material, they pass easily through the emitter, cross over the junction, and combine with holes in the P material (base). For each electron that fills a hole in the P material, another electron will leave the P material (creating a new hole) and enter the positive terminal of the battery.

NPN REVERSE-BIASED JUNCTION. - The second PN junction (base-to-collector), or reverse-biased junction as it is called (fig. 5.1.4), blocks the majority current carriers from crossing the junction. However, there is a very small current, mentioned earlier, that does pass through this junction. This current is called minority current, or reverse current. As you recall, this current was produced by the electron-hole pairs. The minority carriers for the reverse-biased PN junction are the electrons in the P material and the holes in the N material. These minority carriers actually conduct the current for the reverse-biased junction when electrons from the P material enter the N material, and the holes from the N material enter the P material. However, the minority current electrons (as you will see later) play the most important part in the operation of the NPN transistor.

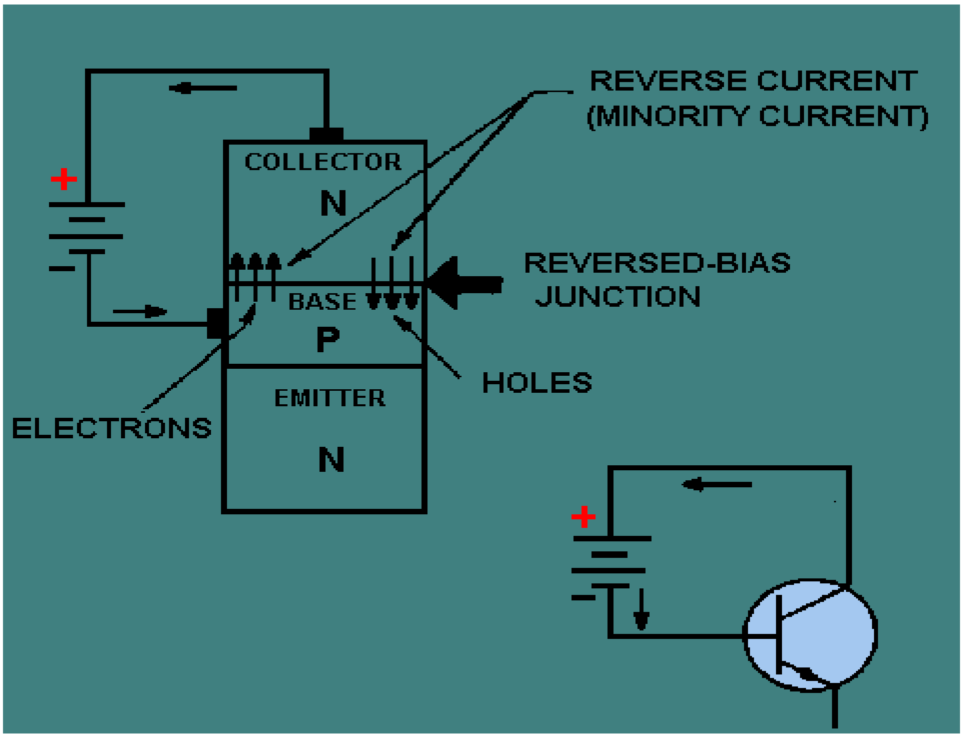


Figure 5.1.4. - The reverse-biased junction in an NPN transistor.

At this point you may wonder why the second PN junction (base-to-collector) is not forward biased like the first PN junction (emitter-to-base). If both junctions were forward biased, the electrons would have a tendency to flow from each end section of the N P N transistor (emitter and collector) to the center P section (base). In essence, we would have two junction diodes possessing a common base, thus eliminating any amplification and defeating the purpose of the transistor. A word of caution is in order at this time. If you should mistakenly bias the second PN junction in the forward direction, the excessive current could develop enough heat to destroy the junctions, making the transistor useless. Therefore, be sure your bias voltage polarities are correct before making any electrical connections.

NPN JUNCTION INTERACTION. - We are now ready to see what happens when we place the two junctions of the NPN transistor in operation at the same time. For a better understanding of just how the two junctions work together, refer to figure 5.1.5 during the discussion.

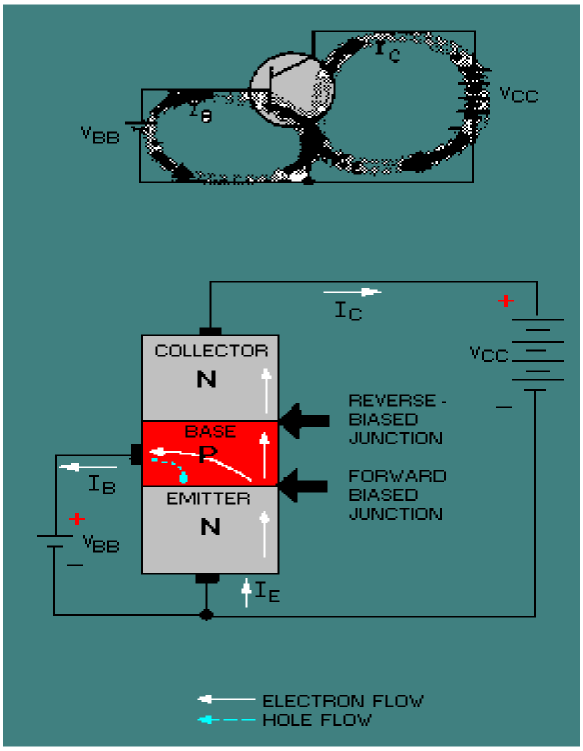


Figure . 5.1.5- NPN transistor operation.

The bias batteries in this figure have been labeled V_{CC} for the collector voltage supply, and V_{BB} for the base voltage supply. Also notice the base supply battery is quite small, as indicated by the number of cells in the battery, usually 1 volt or less. However, the collector supply is generally much higher than the base supply, normally around 6 volts. As you will see later, this difference in supply voltages is necessary to have current flow from the emitter to the collector.

As stated earlier, the current flow in the external circuit is always due to the movement of free electrons. Therefore, electrons flow from the negative terminals of the supply batteries to the N-type emitter. This combined movement of electrons is known as emitter current (I_E). Since electrons are the majority carriers in the N material, they will move through the N material emitter to the emitter-base junction. With this junction forward biased, electrons continue on into the base region. Once the electrons are in the base, which is a P-type material, they become minority carriers. Some of the electrons that move into the base recombine with available holes. For each electron that recombines, another electron moves out through the base lead as base current I_B (creating a new hole for eventual combination) and returns to the base supply battery V Once in the collector, the electrons move easily through the N material and return to the positive terminal of the collector supply battery V_{CC} as collector current (I_C). To further improve on the efficiency of the transistor, the collector is made physically larger than the base for two reasons: (1) to increase the chance of collecting carriers that diffuse to the side as well as directly across the base region, and (2) to enable the collector to handle more heat without damage.

In summary, total current flow in the NPN transistor is through the emitter lead. Therefore, in terms of percentage, I_E is 100 percent. On the other hand, since the base is very thin and lightly doped, a smaller percentage of the total current (emitter current) will flow in the base circuit than in the collector circuit. Usually no more than 2 to 5 percent of the total current is base current (I_B) while the remaining 95 to 98 percent is collector current (I_C). A very basic relationship exists between these two currents:

$I_E = I_B + I_C$

In simple terms this means that the emitter current is separated into base and collector current. Since the amount of current leaving the emitter is solely a function of the emitter-base bias, and because the collector receives most of this current, a small change in emitter-base bias will have a far greater effect on the magnitude of collector current than it will have on base current. In conclusion, the relatively small emitter-base bias controls the relatively large emitter-to-collector current.

5.1.1B. PNP Transistor Operation

The PNP transistor works essentially the same as the NPN transistor. However, since the emitter, base, and collector in the PNP transistor are made of materials that are different from those used in the NPN transistor, different current carriers flow in the PNP unit. The majority current carriers in the PNP transistor are holes. This is in contrast to the NPN transistor where the majority current carriers are electrons. To support this different type of current (hole flow), the bias batteries are reversed for the PNP transistor. A typical bias setup for the PNP transistor is shown in figure 5.1.6.

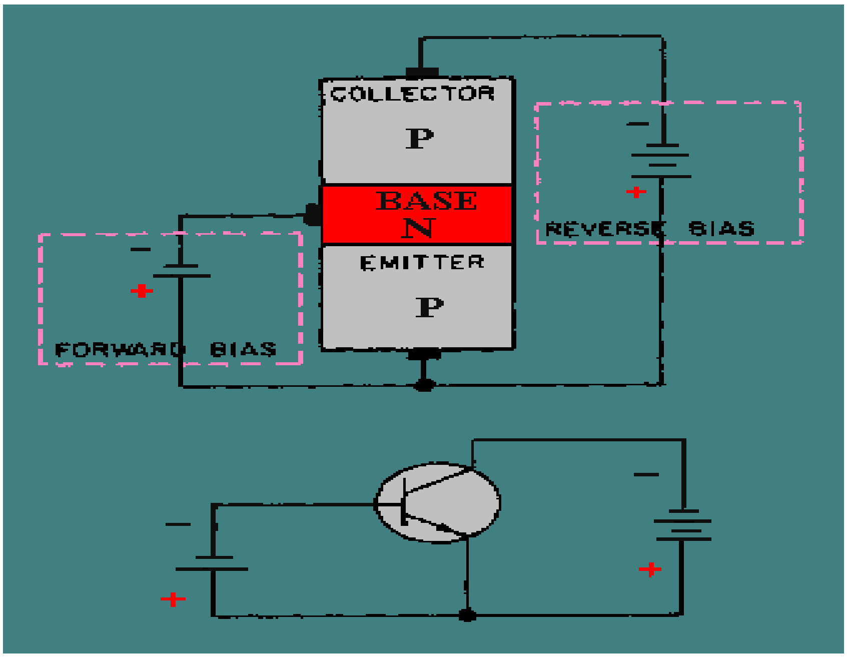


Figure 5.1.6. - A properly biased PNP transistor.

PNP FORWARD-BIASED JUNCTION. - Now let us consider what happens when the emitter-base junction in figure 5.1.7 is forward biased. With the bias setup shown, the positive terminal of the battery repels the emitter holes toward the base, while the negative terminal drives the base electrons toward the emitter. When an emitter hole and a base electron meet, they combine. For each electron that combines with a hole, another electron leaves the negative terminal of the battery, and enters the base. At the same time, an electron leaves the emitter, creating a new hole, and enters the positive terminal of the battery. This movement of electrons into the base and out of the emitter constitutes base current flow (I_B), and the path these electrons take is referred to as the emitter-base circuit.

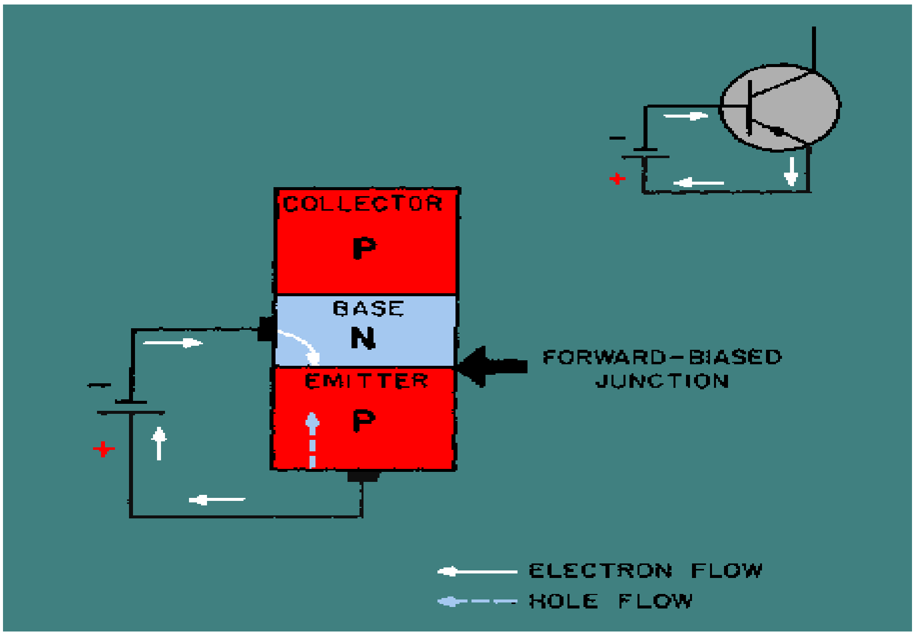


Figure 5.1.7. - The forward-biased junction in a PNP transistor.

PNP REVERSE-BIASED JUNCTION. - In the reverse-biased junction (fig.5.1.8), the negative voltage on the collector and the positive voltage on the base block the majority current carriers from crossing the junction.

However, this same negative collector voltage acts as forward bias for the minority current holes in the base, which cross the junction and enter the collector. The minority current electrons in the collector also sense forward bias-the positive base voltage-and move into the base. The holes in the collector are filled by electrons that flow from the negative terminal of the battery. At the same time the electrons leave the negative terminal of the battery, other electrons in the base break their covalent bonds and enter the positive terminal of the battery. Although there is only minority current flow in the reverse-biased junction, it is still very small because of the limited number of minority current carriers.

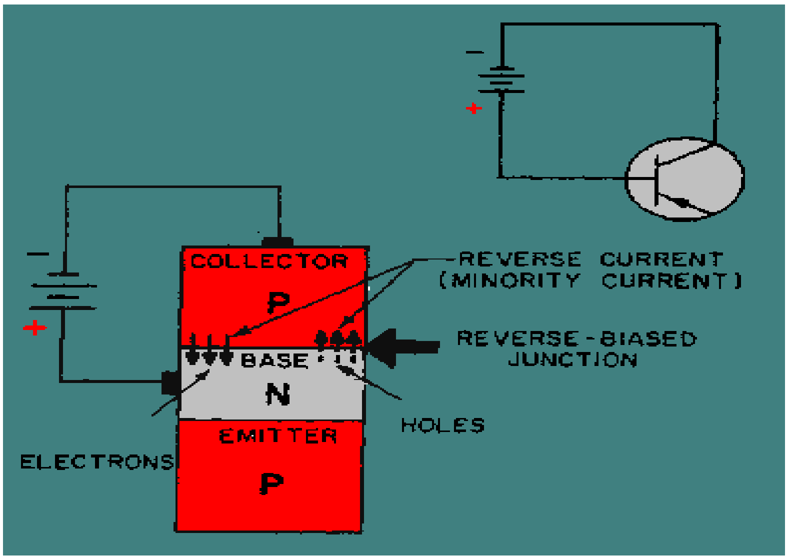


Figure 5.1.8. - The reverse-biased junction in a PNP transistor.

PNP JUNCTION INTERACTION. - The interaction between the forward- and reverse-biased junctions in a PNP transistor is very similar to that in an NPN transistor, except that in the PNP transistor, the majority current carriers are holes. In the PNP transistor shown in figure 5.1.9, the positive voltage on the emitter repels the holes toward the base. Once in the base, the holes combine with base electrons. But again, remember that the base region is made very thin to prevent the recombination of holes with electrons. Therefore, well over 90 percent of the holes that enter the base become attracted to the large negative collector voltage and pass right through the base. However, for each electron and hole that combine in the base region, another electron leaves the negative terminal of the base battery (V_{BB}) and enters the base as base current (I_B). At the same time an electron leaves the negative terminal of the battery, another electron leaves the emitter as IE (creating a new hole) and enters the positive terminal of V_{BB}. Meanwhile, in the collector circuit, electrons from the collector battery (V_{CC}) enter the collector as Ic and combine with the excess holes from the base. For each hole that is neutralized in the collector by an electron, another electron leaves the emitter and starts its way back to the positive terminal of V_{CC}.

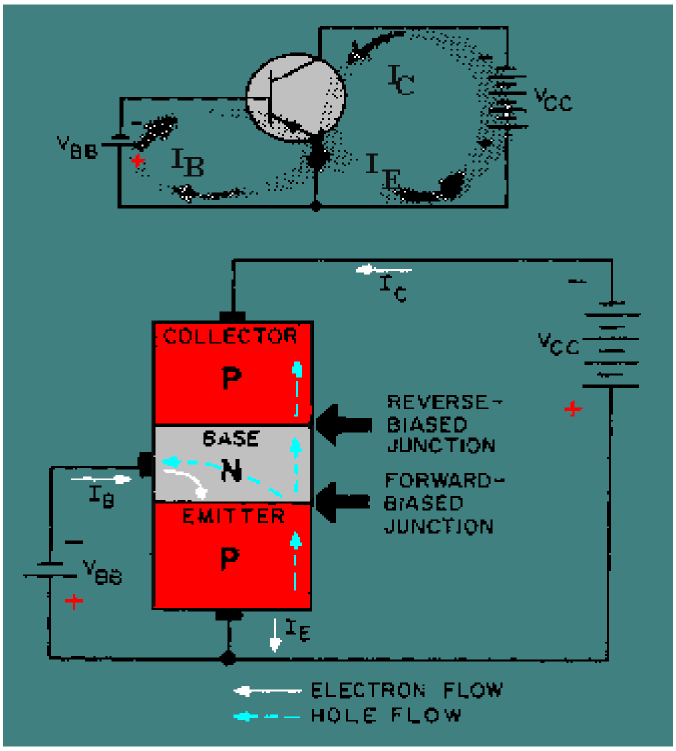
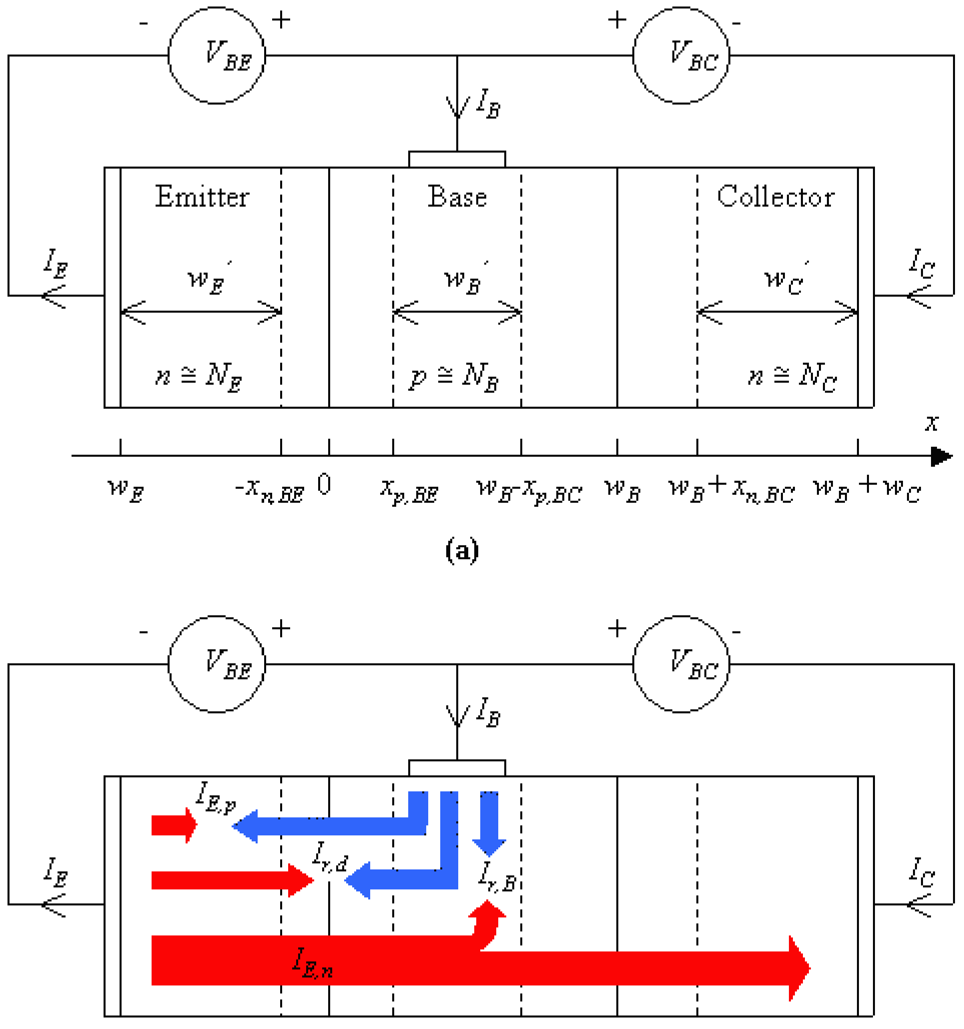


Figure 5.1.9. - PNP transistor operation.

Although current flow in the external circuit of the PNP transistor is opposite in direction to that of the NPN transistor, the majority carriers always flow from the emitter to the collector. This flow of majority carriers also results in the formation of two individual current loops within each transistor. One loop is the base-current path, and the other loop is the collector-current path. The combination of the current in both of these loops ($I_B + I_C$) results in total transistor current (I_E). The most important thing to remember about the two different types of transistors is that the emitter-base voltage of the PNP transistor has the same controlling effect on collector current as that of the NPN transistor. In simple terms, increasing the forward-bias voltage of a transistor reduces the emitter-base junction barrier. This action allows more carriers to reach the collector, causing an increase in current flow from the emitter to the collector and through the external circuit. Conversely, a decrease in the forward-bias voltage reduces collector current.

5.2. Structure and principle of operation:

A bipolar junction transistor consists of two back-to-back p-n junctions, who share a thin common region with width, w_B. Contacts are made to all three regions, the two outer regions called the emitter and collector and the middle region called the base. The structure of an npn bipolar transistor is shown in Figure 5.2.1 (a). The device is called “bipolar” since its operation involves both types of mobile carriers, electrons and holes.



(a)

(b)

Figure 5.2.1.: a) Structure and sign convention of a npn bipolar junction transistor. (b) Electron and hole flow under forward active bias, $V_{BE} > 0$ and $V_{BC} = 0$.

The transistor is a three terminal semiconductor device, consisting of two PN junctions. Usually, the transistor is constructed by "sandwiching" a P-type semiconductor layer between two N-type semiconductor layers as shown in figure 1.

TheBJT'sconstruction gives the appearance of two back-to-back diodes --one from the base to the emitter and another from the base to the collector. The two diode analogy is only a convenient model to understand the states of the transistor. It is neither possible to build a transistor by connecting two diodes back to back, nor is it possible to describe the transistor's operation completely using this model.

The sign convention of the currents and voltage is indicated on Figure 5.2.1(a). The base and collector current are positive if a positive current goes into the base or collector contact. The emitter current is positive for a current coming out of the emitter contact. This also implies that the emitter current, I_E, equals the sum of the base current, I_B, and the collector current, I_C:

$$I_E = I_C + I_B$$

(5.2.1)

The base-emitter voltage and the base-collector voltage are positive if a positive voltage is applied to the base contact relative to the emitter and collector respectively.

The operation of the device is illustrated with Figure 5.2.1 (b). We consider here only the forward active bias mode of operation, obtained by forward biasing the base-emitter junction and reverse biasing the base-collector junction. To simplify the discussion further, we also set V_{CE} = 0. The corresponding energy band diagram is shown in Figure 5.2.2. Electrons diffuse from the emitter into the base and holes diffuse from the base into the emitter. This carrier diffusion is identical to that in a p-n junction. However, what is different is that the electrons can diffuse as minority carriers through the quasi-neutral base. Once the electrons arrive at the base-collector depletion region, they are swept through the depletion layer due to the electric field. These electrons contribute to the collector current. In addition, there are two more currents, the base recombination current, indicated on Figure 5.2.2 by the vertical arrow, and the base-emitter depletion layer recombination current, $I_{r,d}$, (not shown).

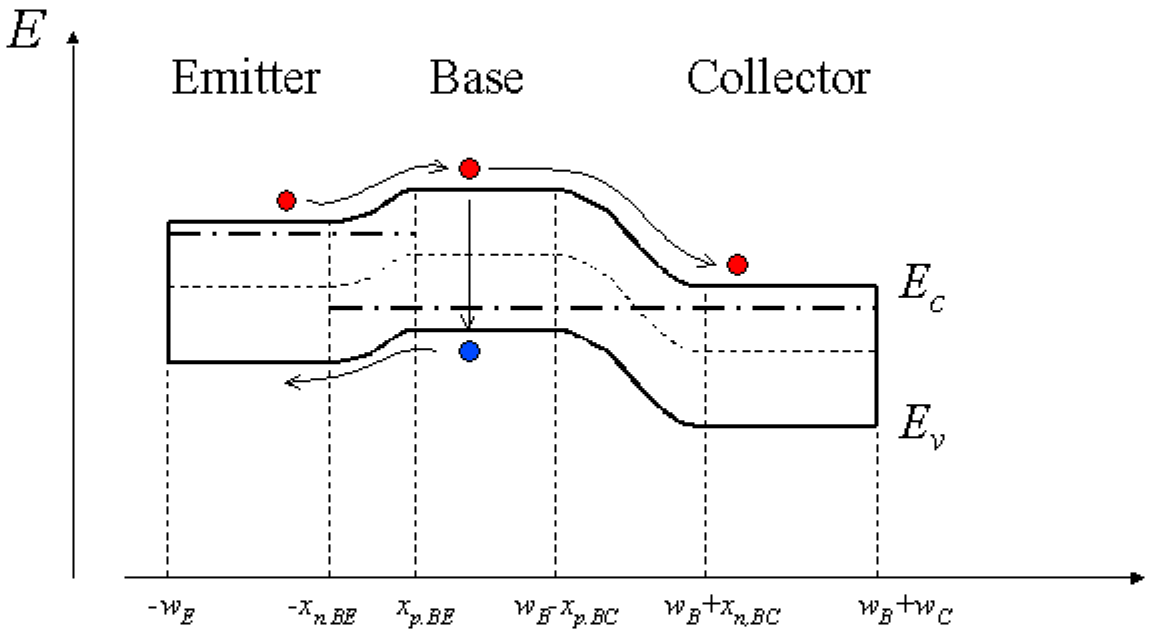


Figure 5.2.2. : Energy band diagram of a bipolar transistor biased in the forward active mode.

The total emitter current is the sum of the electron diffusion current, $I_{E,n}$, the hole diffusion current, $I_{E,p}$ and the base-emitter depletion layer recombination current, $I_{r,d}$.

$$I_E = I_{E,n} + I_{E,p} + I_{r,d}$$
(5.2.2)

The total collector current is the electron diffusion current, $I_{E,n}$, minus the base recombination current, $I_{r,B}$.

$$I_C = I_{E,n} - I_{r,B}$$
(5.2.3)

The base current is the sum of the hole diffusion current, $I_{E,p}$, the base recombination current, $I_{r,B}$ and the base-emitter depletion layer recombination current, $I_{r,d}$.

$$I_B = I_{E,p} + I_{r,B} + I_{r,d}$$
(5.2.4)

The transport factor, α, is defined as the ratio of the collector and emitter current:

$$\alpha = \frac{I_C}{I_E}$$

(5.2.5)

Using Kirchoff's current law and the sign convention shown in Figure 5.2.1(a), we find that the base current equals the difference between the emitter and collector current. The current gain, β, is defined as the ratio of the collector and base current and equals:

$$\beta = \frac{I_C}{I_B} = \frac{\alpha}{1-\alpha}$$

(5.2.6)

This explains how a bipolar junction transistor can provide current amplification. If the collector current is almost equal to the emitter current, the transport factor, α, approaches one. The current gain, β, can therefore become much larger than one.

To facilitate further analysis, we now rewrite the transport factor, α, as the product of the emitter efficiency, γ_E, the base transport factor, α_T, and the depletion layer recombination factor, δ_r.

$$\alpha = \alpha_T \, \gamma_E \, \delta_r$$

(5.2.7)

The emitter efficiency, γ_E, is defined as the ratio of the electron current in the emitter, $I_{E,n}$, to the sum of the electron and hole current diffusing across the base-emitter junction, $I_{E,n} + I_{E,p}$.

$$\gamma_E = \frac{I_{E.n}}{I_{E,n} + I_{E,p}}$$

(5.2.8)

The base transport factor, α_T, equals the ratio of the current due to electrons injected in the collector, to the current due to electrons injected in the base.

$$\alpha_T = \frac{I_{E,n} - I_{r,B}}{I_E,n}$$

(5.2.9)

Recombination in the depletion-region of the base-emitter junction further reduces the current gain, as it increases the emitter current without increasing the collector current. The depletion layer recombination factor, δ_r, equals the ratio of the current due to electron and hole diffusion across the base-emitter junction to the total emitter current:

$$\delta_r = \frac{I_E - I_{r,d}}{I_E}$$

(5.2.10)

Example

A bipolar transistor with an emitter current of 1 mA has an emitter efficiency of 0.99, a base transport factor of 0.995 and a depletion layer recombination factor of 0.998. Calculate the base current, the collector current, the transport factor and the current gain of the transistor.

Solution

The transport factor and current gain are:

$$\alpha = \gamma_E \alpha_T \delta_r = 0.99 \times 0.995 \times 0.998 = 0.983$$

and

$$\beta = \frac{\alpha}{1-\alpha} = 58.1$$

The collector current then equals

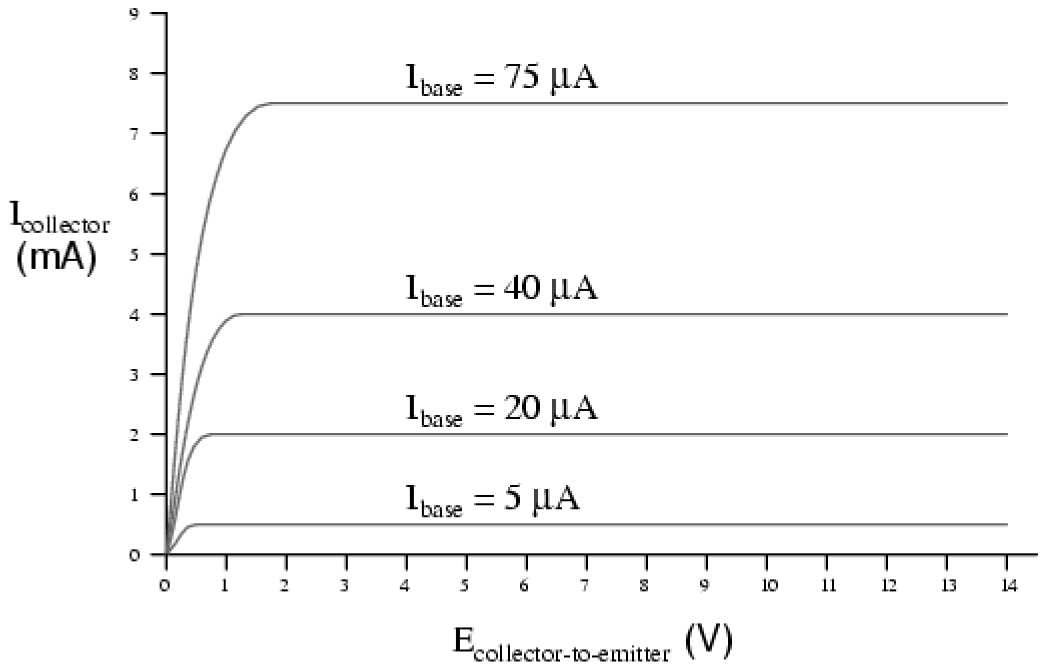
$$I_C = \alpha I_E = 0.983\,\text{mA}$$

And the base current is obtained from:

$$I_B = I_E - I_C = 17\,\mu\text{A}$$

I-V charactesistics of a Transistor:

Often it is useful to superimpose several collector current/voltage graphs for different base currents on the same graph. A collection of curves like this -- one curve plotted for each distinct level of base current -- for a particular transistor is called the transistor's *characteristic curves*:



Each curve on the graph reflects the collector current of the transistor, plotted over a range of collector-to-emitter voltages, for a given amount of base current. Since a transistor tends to act as a current regulator, limiting collector current to a proportion set by the base current, it is useful to express this proportion as a standard transistor performance measure. Specifically, the ratio of collector current to base current is known as the *Beta* ratio (symbolized by the Greek letter β):

$$\beta = \frac{I_{collector}}{I_{base}}$$

β is also known as h_{fe}

Sometimes the β ratio is designated as "h_{fe}," a label used in a branch of mathematical semiconductor analysis known as "hybrid parameters" which strives to achieve very precise predictions of transistor performance with detailed equations. Hybrid parameter variables are many, but they are all labeled with the general letter "h" and a specific subscript. The variable "h_{fe}" is just another (standardized) way of expressing the ratio of collector current to base current, and is interchangeable with "β." Like all ratios, β is unitless.

β for any transistor is determined by its design: it cannot be altered after manufacture. However, there are so many physical variables impacting β that it is rare to have two transistors of the same design exactly match. If a circuit design relies on equal β ratios between multiple transistors, "matched sets" of transistors may be purchased at extra cost. However, it is generally considered bad design practice to engineer circuits with such dependencies.

5.2.2: Relation between β and α :

$$\beta = \frac{I_C}{I_B}, \alpha = \frac{I_C}{I_E}$$

We know

$I_E=I_B+I_C$

So, $\beta = \frac{I_C}{I_E - I_C} = \frac{I_C / I_E}{(I_E / I_E) - (I_C / I_E)} = \frac{\alpha}{1 - \alpha}$

Ideal transistor model

The ideal transistor model is based on the ideal p-n diode model and provides a first-order calculation of the dc parameters of a bipolar junction transistor. To further simplify this model, we will assume that all quasi-neutral regions in the device are much smaller than the minority-carrier diffusion lengths in these regions, so that the "short" diode expressions apply. The use of the ideal p-n diode model implies that no recombination within the depletion regions is taken into account. The discussion of the ideal transistor starts with a discussion of the forward active mode of operation, followed by a general description of the four different bias modes, the corresponding Ebers-Moll model and a calculation of the collector-emitter voltage when the device is biased in saturation.

General bias modes of a bipolar transistor

While the forward active mode of operation is the most useful bias mode when using a bipolar junction transistor as an amplifier, one cannot ignore the other bias modes especially when using the device as a digital switch. All possible bias modes are illustrated with Figure 5.3.2. They are the forward active mode of operation, the reverse active mode of operation, the saturation mode and the cut-off mode.

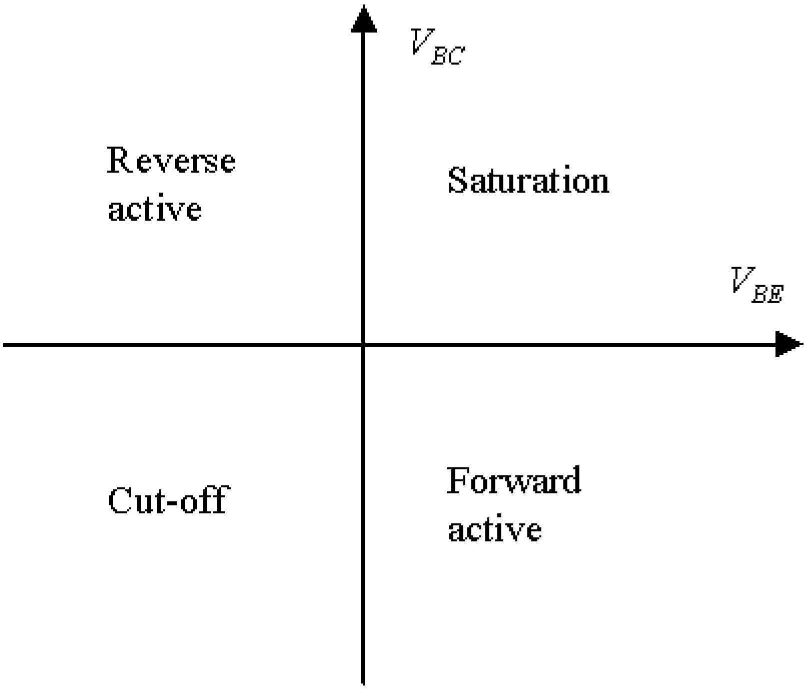


Figure 5.3.2.: Possible bias modes of operation of a bipolar junction transistor.

Active Mode Operation:

Forward Active Mode:

The forward active mode is the one where we forward bias the base-emitter junction, $V_{\mathrm{BE}} > 0$ and reverse bias the base-collector junction, $V_{\mathrm{BC}} < 0$.

Reverse Active Mode:

Finally, there is the reverse active mode of operation. In the reverse active mode, we reverse the function of the emitter and the collector. We reverse bias the base-emitter junction and forward bias the base-collector junction, or $V_{\mathrm{BE}} < 0$ and $V_{\mathrm{BC}} > 0$. Most transistors, however, have poor emitter efficiency under reverse active bias since the collector doping density is typically much less than the base doping density to ensure high base-collector breakdown voltages. In addition, the collector-base area is typically larger than the emitter-base area, so that even fewer electrons make it from the collector into the emitter.

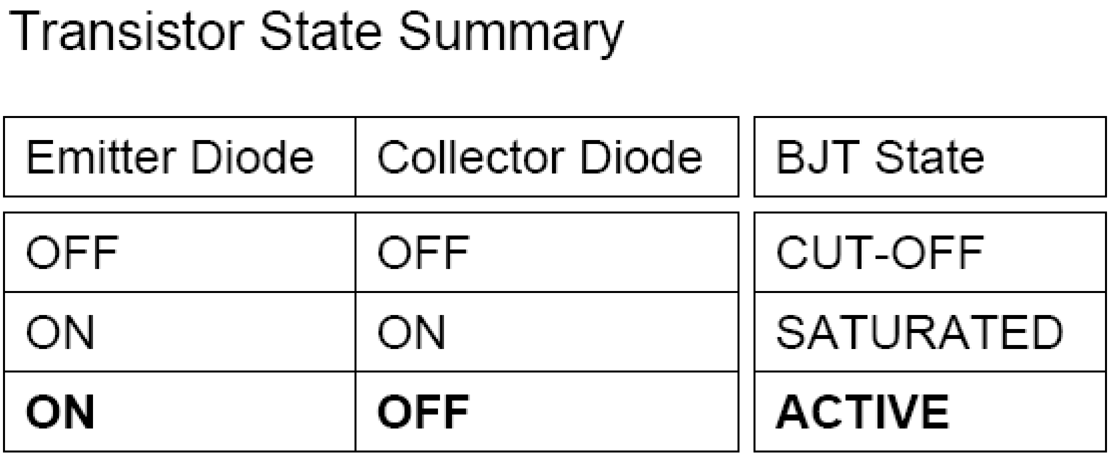
Cut-Off Mode:

In the cut-off mode, both junctions are reversed biased, $V_{\mathrm{BE}} < 0$ and $V_{\mathrm{BC}} < 0$, so that very little current goes through the device. This corresponds to the “off” state of the device.

Saturation Mode:

In the saturation mode, both junctions are forward biased, $V_{\mathrm{BE}} > 0$ and $V_{\mathrm{CB}} > 0$. This corresponds to the low resistance “on” state of the transistor.

Saturation also implies that a large amount of minority carrier charge is accumulated in the base region. As a transistor is switched from saturation to cut-off, this charge initially remains in the base and a collector current will remain until this charge is removed by recombination. This causes an additional delay before the transistor is turned off. Since the carrier lifetime can be significantly longer than the base transit time, the turn-off delay causes a large and undesirable asymmetry between turn-on and turn-off time. Saturation is therefore avoided in high-speed bipolar logic circuits. Two techniques are used to reduce the turn-off delay: 1) adding a Schottky diode in parallel to the base-collector junction and 2) using an emitter-coupled circuit configuration. Both approaches avoid biasing the transistor in the saturation mode. The Schottky diode clamps the base-collector voltage at a value, which is slightly lower than the turn-on voltage of the base-collector diode. An emitter-coupled circuit is biased with a current source, which can be designed such that the collector voltage cannot be less than the base voltage.



Transistor State Summary

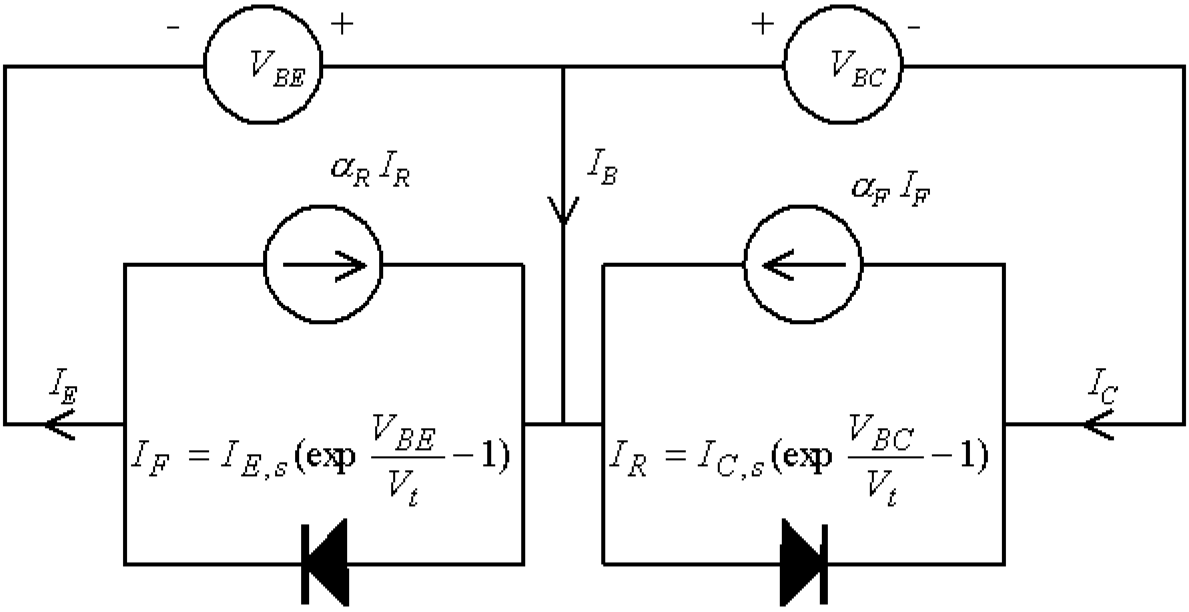
Emitter Diode	Collector Diode	BJT State
OFF	OFF	CUT-OFF
ON	ON	SATURATED
ON	**OFF**	**ACTIVE**

The Ebers-Moll model describes all of these bias modes.

The Ebers-Moll model

The Ebers-Moll model is an ideal model for a bipolar transistor, which can be used, in the forward active mode of operation, in the reverse active mode, in saturation and in cut-off. This model is the predecessor of today's computer simulation models and contains only the "ideal" diode currents.

The model contains two diodes and two current sources as shown in Figure 5.3.3. The two diodes represent the base-emitter and base-collector diodes. The current sources quantify the transport of minority carriers through the base region. These current sources depend on the current through each diode. The parameters $I_{E,s}$, $I_{C,s}$, α_F and α_R are the saturation currents of the base-emitter and base collector diode and the forward and reverse transport factors.



-
+
V_{BE}
+
-
V_{BC}
$\alpha_R I_R$
I_B
$\alpha_F I_F$
I_E
I_C
$I_F = I_{E,s}(\exp \frac{V_{BE}}{V_t} - 1)$
$I_R = I_{C,s}(\exp \frac{V_{BC}}{V_t} - 1)$

Figure 6.3.3 : Equivalent circuit for the Ebers-Moll model of an npn bipolar junction transistor

Using the parameters identified in Figure 5.3.3, we can relate the emitter, base and collector current to the forward and reverse currents and transport factors, yielding:

$$I_E = I_F - \alpha_R I_R$$

(5.3.1)

$$I_B = (1 - \alpha_F) I_F + (1 - \alpha_R) I_R$$

(5.3.2)

$$I_C = -I_R + \alpha_F I_F$$

(5.3.3)

The Ebers-Moll parameters are related by the following equation:

$$I_{E,s} \alpha_F = I_{C,s} \alpha_R$$

(5.3.4)

This relation ship is also referred as the reciprocity relation and can be derived by examining the minority carrier current through the base. For the specific case where the base-emitter and base-collector voltage are the same and the base doping is uniform, there can be no minority carrier diffusion in the base so that:

$$I_F (V_{BE}) \alpha_F = I_R (V_{BC} = V_{BE}) \alpha_R$$

(5.3.5)

from which the reciprocity relation is obtained.

The forward- and reverse-bias transport factors are obtained by measuring the current gain in the forward active and reverse active mode of operation. The saturation currents $I_{E,s}$ and $I_{C,s}$ are obtained by measuring the base-emitter (base-collector) diode saturation current while shorting the base-collector (base-emitter) diode.

Saturation(Described by The Ebers-Moll model)

In the low resistance "on" state of a bipolar transistor, one finds that the voltage between the collector and emitter is less than the forward bias voltage of the base-emitter junction. Typically the "on" state voltage of a silicon BJT is 100 mV and the forward bias voltage is 700 mV. Therefore, the base-collector junction is also forward biased. Using the Ebers-Moll model, we can calculate the "on" voltage from:

$$V_{CE,sat} = V_{BE} - V_{BC} = V_t \ln \left\{ \frac{I_F}{I_R} \frac{I_{C,s}}{I_{E,s}} \right\}$$

(5.3.6)

and using equations (5.3.3), (5.3.4) and the reciprocity relation (5.3.5), one obtains:

$$V_{CE,sat} = V_t \ln\left\{\frac{1+\frac{I_C}{I_B}(1-\alpha_R)}{\alpha_R\left[1-\frac{I_C}{I_B}\frac{(1-\alpha_F)}{\alpha_F}\right]}\right\}$$

(5.3.7)

Example
Calculate the saturation voltage of a bipolar transistor biased with a base current of 1 mA and a collector current of 10 mA. Use $\alpha_R = 0.993$ and $\alpha_F = 0.2$.

Solution

The saturation voltage equals:

$$V_{CE,sat} = V_t \ln\left\{\frac{1+\frac{I_C}{I_B}(1-\alpha_R)}{\alpha_R\left[1-\frac{I_C}{I_B}\frac{(1-\alpha_F)}{\alpha_F}\right]}\right\} = 0.1\,\mathrm{V}$$

Base spreading resistance and emitter current crowding

Large area bipolar transistors can have a very non-uniform current distribution due to the resistance of the base layer. Since the base current is applied through the thin base layer, there can be a significant series resistance in large devices. This resistance causes a voltage variation across the base region. This voltage variation then causes a variation of the emitter current density, especially since the emitter current density depends exponentially on the local base-emitter voltage. This effect is minimal in the center of the emitter-base diode and strongly increases toward the edges. In extreme cases, this effect causes the emitter current to occur only at the very edges of the emitter-base diode. The parameters involved include the sheet resistance of the base layer, the emitter current density and the current gain in the device. The characteristic length, $\lambda_{spreading}$, can be obtained from a distributed model similar to that of a metal contact to a thin semiconductor layer.

$$\lambda_{spreading} = \sqrt{\frac{r_\pi\, Area}{R_{s,B}}} = \sqrt{\frac{V_t\, \beta}{J_E R_{s,B}}}$$

(5.4.1)

Where r_π is the small signal base resistance, $R_{s,B}$ is the sheet resistance of the base and J_E is the emitter current density. This analysis is only valid if the emitter current density is close to uniform. The emitter current density in a BJT can only be consider close to uniform if the emitter stripe width is less that the characteristic length for a BJT with a one-sided base contact or less that twice the characteristics length for a BJT with a double sided base contact or:

$$W_{s,E} \le 2\,\lambda_{spreading}$$

(5.4.2)

The corresponding value of the base resistance for a uniform emitter current distribution equals:

$$R_B = \frac{1}{3} R_s \frac{W_{s,E}}{L_{s,E}}$$

(5.4.3)

for a one-sided base contact and

$$R_B = \frac{1}{12} R_s \frac{W_{s,E}}{L_{s,E}}$$

(5.4.4)

for a double-sided base contact, which effectively has the resistance of two sections with half the emitter stripe width connected in parallel. A series of narrow emitter fingers with alternating base contacts is therefore typically used in large area power devices, resulting in the characteristic interdigitated structure.

Temperature dependent effects in bipolar transistors

The temperature dependence of bipolar transistors depends on a multitude of parameters affecting the bipolar transistor characteristics in different ways.

First we will discuss the temperature dependence of the current gain. Since the current gain depends on both the emitter efficiency and base transport factor, we will discuss these separately.

The emitter efficiency depends on the ratio of the carrier density, diffusion constant and width of the emitter and base. As a result, it is not expected to be very temperature dependent. The carrier densities are linked to the doping densities. Barring incomplete ionization, which can be very temperature dependent, the carrier densities are independent of temperature as long as the intrinsic carrier density does not exceed the doping density in either region. The width is very unlikely to be temperature dependent and therefore also the ratio of the emitter and base width. The ratio of the mobility is expected to be somewhat temperature dependent due to the different temperature dependence of the mobility in n-type and p-type material.

The base transport is more likely to be temperature dependent since it depends on the product of the diffusion constant and carrier lifetime. The diffusion constant in turn equals the product of the thermal voltage and the minority carrier mobility in the base. The recombination lifetime depends on the thermal velocity. The result is therefore moderately dependent on temperature. Typically the base transport reduces with temperature, primarily because the mobility and recombination lifetime are reduced with increasing temperature. Occasionally the transport factor initially increases with temperature, but then reduces again.

The Basic Transistor Amplifier

In the preceding pages we explained the internal workings of the transistor and introduced new terms, such as emitter, base, and collector. Since you should be familiar by now with all of the new terms mentioned earlier and with the internal operation of the transistor, we will move on to the basic transistor amplifier.

To understand the overall operation of the transistor amplifier, you must only consider the current in and out of the transistor and through the various components in the circuit. Therefore, from this point on, only the schematic symbol for the transistor will be used in the illustrations, and rather than thinking about majority and minority carriers, we will now start thinking in terms of emitter, base, and collector current.

Before going into the basic transistor amplifier, there are two terms you should be familiar with: AMPLIFICATION and AMPLIFIER. Amplification is the process of increasing the strength of a SIGNAL. A signal is just a general term used to refer to any particular current, voltage, or power in a circuit. An amplifier is the device that provides amplification (the increase in current, voltage, or power of a signal) without appreciably altering the original signal.

Transistors are frequently used as amplifiers. Some transistor circuits are CURRENT amplifiers, with a small load resistance; other circuits are designed for VOLTAGE amplification and have a high load resistance; others amplify POWER.

Now take a look at the NPN version of the basic transistor amplifier in figure 6.6.1 and let's see just how it works.

So far in this discussion, a separate battery has been used to provide the necessary forward-bias voltage. Although a separate battery has been used in the past for convenience, it is not practical to use a battery for emitter-base bias. For instance, it would take a battery slightly over .2 volts to properly forward bias a germanium transistor, while a similar silicon transistor would require a voltage slightly over .6 volts. However, common batteries do not have such voltage values. Also, since bias voltages are quite critical and must be held within a few tenths of one volt, it is easier to work with bias currents flowing through resistors of high ohmic values than with batteries.

By inserting one or more resistors in a circuit, different methods of biasing may be achieved and the emitter-base battery eliminated. In addition to eliminating the battery, some of these biasing methods compensate for slight variations in transistor characteristics and changes in transistor conduction resulting from temperature irregularities. Notice in figure 6.6.1 that the emitter-base battery has been eliminated and the bias resistor R_B has been inserted between the collector and the base. Resistor R_B provides the necessary forward bias for the emitter-base junction. Current flows in the emitter-base bias circuit from ground to the emitter, out the base lead, and through R_B to V_{CC}. Since the current in the base circuit is very small (a few hundred microamperes) and the forward resistance of the transistor is low, only a few tenths of a volt of positive bias will be felt on the base of the transistor. However, this is enough voltage on the base, along with ground on the emitter and the large positive voltage on the collector, to properly bias the transistor.

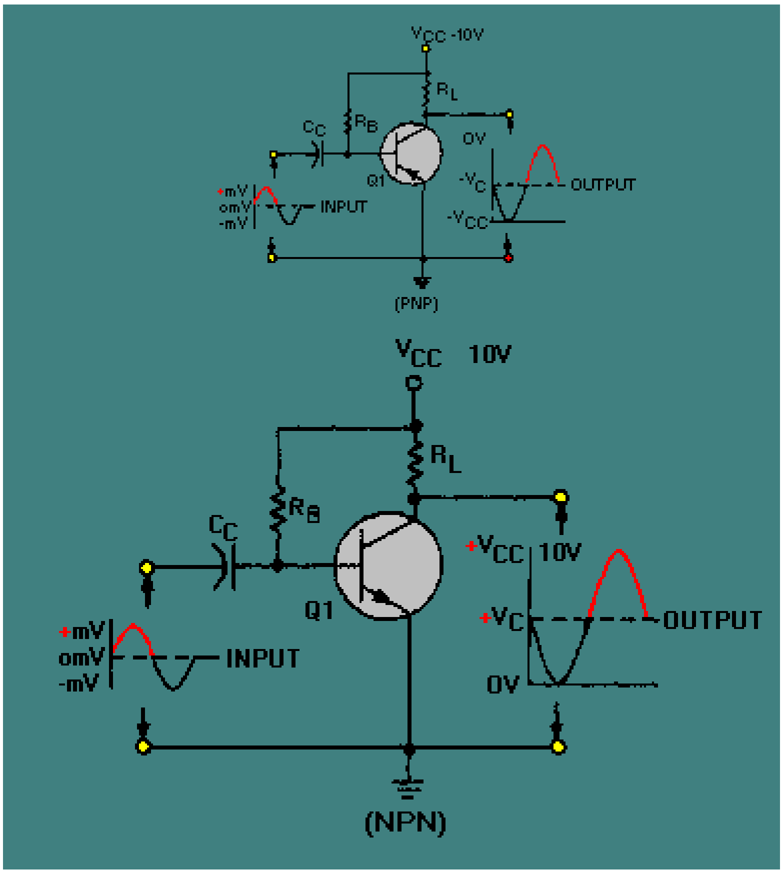


Figure 5.6.1. - The basic transistor amplifier.

With Q1 properly biased, direct current flows continuously, with or without an input signal, throughout the entire circuit. The direct current flowing through the circuit develops more than just base bias; it also develops the collector voltage (V_C) as it flows through Q1 and R_L. Notice the collector voltage on the output graph. Since it is present in the circuit without an input signal, the output signal starts at the V_C level and either increases or decreases. These dc voltages and currents that exist in the circuit before the application of a signal are known as QUIESCENT voltages and currents (the quiescent state of the circuit).

Resistor R_L, the collector load resistor, is placed in the circuit to keep the full effect of the collector supply voltage off the collector. This permits the collector voltage (V_C) to change with an input signal, which in turn allows the transistor to amplify voltage. Without R_L in the circuit, the voltage on the collector would always be equal to V_{CC}.

The coupling capacitor (C_C) is another new addition to the transistor circuit. It is used to pass the ac input signal and block the dc voltage from the preceding circuit. This prevents dc in the circuitry on the left of the coupling capacitor from affecting the bias on Q1. The coupling capacitor also blocks the bias of Q1 from reaching the input signal source.

The input to the amplifier is a sine wave that varies a few millivolts above and below zero. It is introduced into the circuit by the coupling capacitor and is applied between the base and emitter. As the input signal goes positive, the voltage across the emitter-base junction becomes more positive. This in effect increases forward bias, which causes base current to increase at the same rate as that of the input sine wave. Emitter and collector currents also increase but much more than the base current. With an increase in collector current, more voltage is developed across R_L. Since the voltage across R_L and the voltage across Q1 (collector to emitter) must add up to V_{CC}, an increase in voltage across R_L results in an equal decrease in voltage across Q1. Therefore, the output voltage from the amplifier, taken at the collector of Q1 with respect to the emitter, is a negative alternation of voltage that is larger than the input, but has the same sine wave characteristics.

During the negative alternation of the input, the input signal opposes the forward bias. This action decreases base current, which results in a decrease in both emitter and collector currents. The decrease in current through R_L decreases its voltage drop and causes the voltage across the transistor to rise along with the output voltage. Therefore, the output for the negative alternation of the input is a positive alternation of voltage that is larger than the input but has the same sine wave characteristics.

By examining both input and output signals for one complete alternation of the input, we can see that the output of the amplifier is an exact reproduction of the input except for the reversal in polarity and the increased amplitude (a few millivolts as compared to a few volts).

The PNP version of this amplifier is shown in the upper part of the figure. The primary difference between the NPN and PNP amplifier is the polarity of the source voltage. With a negative V_{CC}, the PNP base voltage is slightly negative with respect to ground, which provides the necessary forward bias condition between the emitter and base.

When the PNP input signal goes positive, it opposes the forward bias of the transistor. This action cancels some of the negative voltage across the emitter-base junction, which reduces the current through the transistor. Therefore, the voltage across the load resistor decreases, and the voltage across the transistor increases. Since V_{CC} is negative, the voltage on the collector (V_C) goes in a negative direction (as shown on the output graph) toward $-V_{CC}$ (for example, from -5 volts to -7 volts). Thus, the output is a negative alternation of voltage that varies at the same rate as the sine wave input, but it is opposite in polarity and has a much larger amplitude .

During the negative alternation of the input signal, the transistor current increases because the input voltage aids the forward bias. Therefore, the voltage across R_L increases, and consequently, the voltage across the transistor decreases or goes in a positive direction (for example: from -5 volts to -3 volts). This action results in a positive output voltage, which has the same characteristics as the input except that it has been amplified and the polarity is reversed.

In summary, the input signals in the preceding circuits were amplified because the small change in base current caused a large change in collector current. And, by placing resistor R_L in series with the collector, voltage amplification was achieved

Transistor Configurations

A transistor may be connected in any one of three basic configurations (fig. 5.7.1): common emitter (CE), common base (CB), and common collector (CC). The term common is used to denote the element that is common to both input and output circuits. Because the common element is often grounded, these configurations are frequently referred to as grounded emitter, grounded base, and grounded collector.

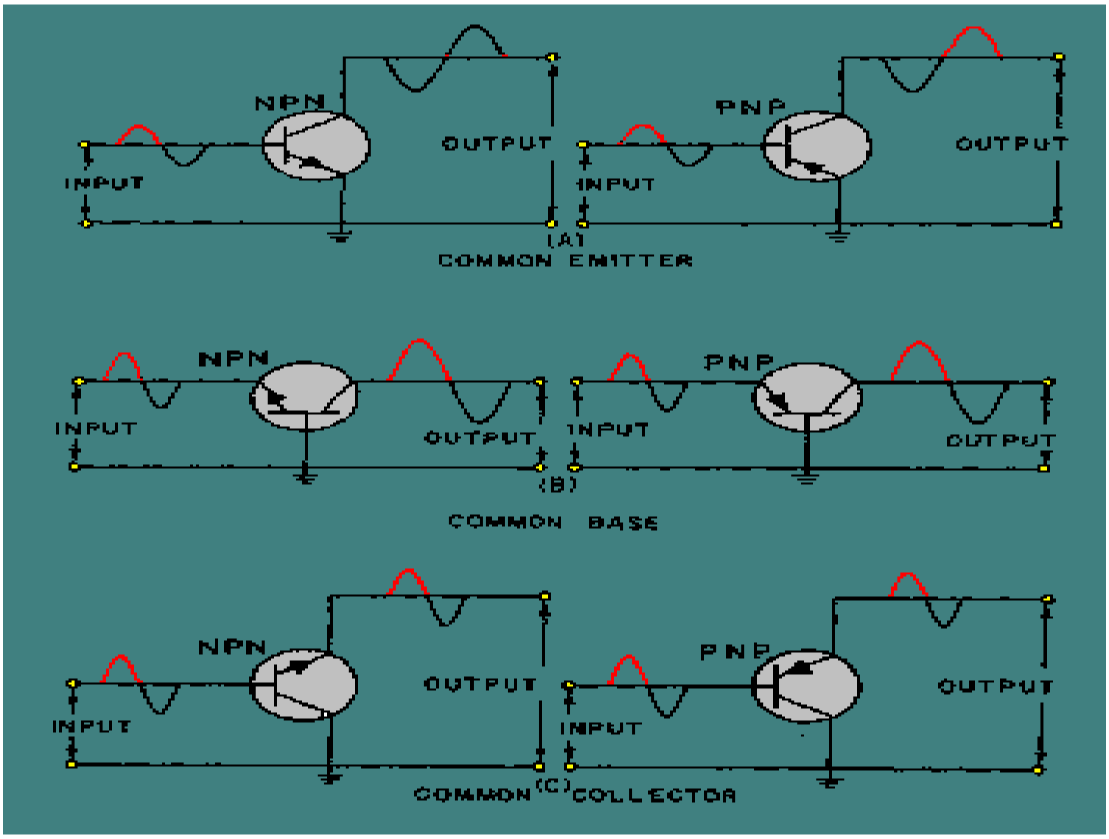


Figure 5.7.1. - Transistor configurations.

Each configuration, as you will see later, has particular characteristics that make it suitable for specific applications. An easy way to identify a specific transistor configuration is to follow three simple steps:

- Identify the element (emitter, base, or collector) to which the input signal is applied.
- Identify the element (emitter, base, or collector) from which the output signal is taken.
- The remaining element is the common element, and gives the configuration its name.

Therefore, by applying these three simple steps to the circuit in figure 6.6.1, we can conclude that this circuit is more than just a basic transistor amplifier. It is a common-emitter amplifier.

Common Base

The common-base configuration (CB) shown in figure 5.7.1, view B is mainly used for impedance matching, since it has a low input resistance (30 ohms-160 ohms) and a high output resistance (250 kilohms-550 kilohms). However, two factors limit its usefulness in some circuit applications: (1) its low input resistance and (2) its current gain of less than 1. Since the CB configuration will give voltage amplification, there are some additional applications, which require both a low-input resistance and voltage amplification, that could use a circuit configuration of this type; for example, some microphone amplifiers.

In the common-base configuration, the input signal is applied to the emitter, the output is taken from the collector, and the base is the element common to both input and output. Since the input is applied to the emitter, it causes the emitter-base junction to react in the same manner as it did in the common-emitter circuit. For example, an input that aids the bias will increase transistor current, and one that opposes the bias will decrease transistor current.

Unlike the common-emitter circuit, the input and output signals in the common-base circuit are in phase. To illustrate this point, assume the input to the PNP version of the common-base circuit in figure 5.7.1 view B is positive. The signal adds to the forward bias, since it is applied to the emitter, causing the collector current to increase. This increase in Ic results in a greater voltage drop across the load resistor R_L (not shown), thus lowering the collector voltage V_C. The collector voltage, in becoming less negative, is swinging in a positive direction, and is therefore in phase with the incoming positive signal.

The current gain in the common-base circuit is calculated in a method similar to that of the common emitter except that the input current is I_E not I_B and the term ALPHA (a) is used in place of beta for gain. Alpha is the relationship of collector current (output current) to emitter current (input current). Alpha is calculated using the formula:

$$\alpha = \frac{\Delta I_C}{\Delta I_E}$$

For example, if the input current (I_E) in a common base changes from 1 mA to 3 mA and the output current (I_C) changes from 1 mA to 2.8 mA, the current gain (a) will be 0.90 or:

$$\alpha = \frac{\Delta I_C}{\Delta I_E} = \frac{18 \times 10^{-3}}{2 \times 10^{-3}} = 0.90$$

This is a current gain of less than 1.

Since part of the emitter current flows into the base and does not appear as collector current, collector current will always be less than the emitter current that causes it. (Remember, $I_E = I_B + I_C$) Therefore, ALPHA is ALWAYS LESS THAN ONE FOR A COMMON-BASE CONFIGURATION.

Another term for "a" is h_{fb}. These terms (and h_{fb}) are equivalent and may be used interchangeably. The meaning for the term h_{fb} is derived in the same manner as the term h_{fe} mentioned earlier, except that the last letter "e" has been replaced with "b" to stand for common- base configuration.

Many transistor manuals and data sheets only list transistor current gain characteristics in terms of b or h_{fe}. To find alpha (a) when given beta (b), use the following formula to convert b to a for use with the common-base configuration:

$$\alpha = \frac{\beta}{\beta + 1}$$

To calculate the other gains (voltage and power) in the common-base configuration when the current gain (a) is known, follow the procedures described earlier under the common-emitter section.

It is seen that total collector current is actually the sum of two components:

(i) Current produced by normal transistor action. Its value is αI_E and is due to majority carriers.

(ii) Temperature dependent leakage current I_{CO} due to minority carriers.

$$I_C = \alpha I_E + I_{CO}$$

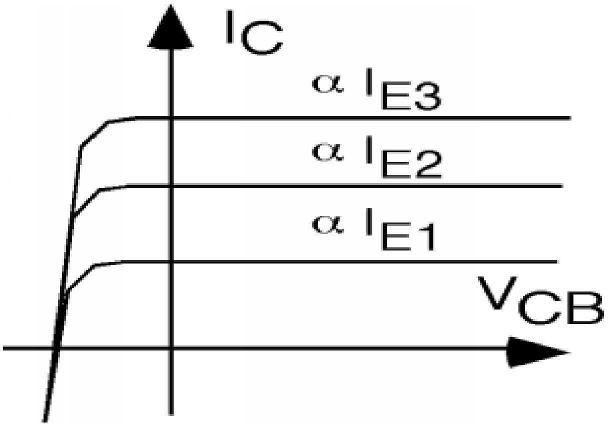
Again I_E=I_B+I_C

$$I_C = \alpha(I_C + I_B) + I_{CO}$$

$$I_C = \frac{\alpha I_B}{1-\alpha} + \frac{I_{CO}}{1-\alpha}$$

$$I_B = (1-\alpha)I_E + I_{CO}$$

I-V Characteristics of Common Base Configuration:



The only drawback with what we have so far is that except in some specialized high-frequency circuits, the bipolar transistor is very rarely used in the common base configuration.

Common Emitter

The common-emitter configuration (CE) shown in figure 5.7.1 view A is the arrangement most frequently used in practical amplifier circuits, since it provides good voltage, current, and power gain. The common emitter also has a somewhat low input resistance (500 ohms-1500 ohms), because the input is applied to the forward-biased junction, and a moderately high output resistance (30 kilohms-50 kilohms or more), because the output is taken off the reverse-biased junction. Since the input signal is applied to the base-emitter circuit and the output is taken from the collector-emitter circuit, the emitter is the element common to both input and output.

When a transistor is connected in a common-emitter configuration, the input signal is injected between the base and emitter, which is a low resistance, low-current circuit. As the input signal swings positive, it also causes the base to swing positive with respect to the emitter. This action decreases forward bias which reduces collector current (I_C) and increases collector voltage (making V_C more negative). During the negative alternation of the input signal, the base is driven more negative with respect to the emitter. This increases forward bias and allows more current carriers to be released from the emitter, which results in an increase in collector current and a decrease in collector voltage (making V_C less negative or swing in a positive direction). The collector current that flows through the high resistance reverse-biased junction also flows through a high resistance load (not shown), resulting in a high level of amplification.

Since the input signal to the common emitter goes positive when the output goes negative, the two signals (input and output) are 180 degrees out of phase. The common-emitter circuit is the only configuration that provides a phase reversal.

The common-emitter is the most popular of the three transistor configurations because it has the best combination of current and voltage gain. The term *GAIN* is used to describe the amplification capabilities of the amplifier. It is basically a ratio of output versus input. Each transistor configuration gives a different value of gain even though the same transistor is used. The transistor configuration used is a matter of design consideration. However, as a technician you will become interested in this output versus input ratio (gain) to determine whether or not the transistor is working properly in the circuit.

In the common-emitter circuit of an n-p-n transistor whose base lead is open. It is found that despite $I_B=0$, there is a leakage current from collector to emitter. It is called $I_{CEO.}$

Taking the leakage current into account, the current distribution through a CE circuit becomes

$$I_C = \beta I_B + I_{CEO} = \beta I_B + (1+\beta) I_{CO} = \beta I_B + \frac{I_{CO}}{1-\alpha}$$

Now

$$I_C = \frac{\alpha I_B}{1-\alpha} + \frac{I_{CO}}{1-\alpha} \text{ because } \beta = \frac{\alpha}{1-\alpha}$$

Again $\beta I_B = \alpha I_E$

$$I_C = \alpha I_E + I_{CEO}$$

$$I_B = (1-\alpha) I_E - I_{CEO}$$

The current gain in the common-emitter circuit is called BETA (b). Beta is the relationship of collector current (output current) to base current (input current). To calculate beta, use the following formula:

$$\beta = \frac{\Delta I_C}{\Delta I_B}$$

(Δ is the Greek letter delta, it is used to indicate a small change)

For example, if the input current (I_B) in a common emitter changes from 75 mA to 100 mA and the output current (I_C) changes from 1.5 mA to 2.6 mA, the current gain (b) will be 44.

$$\beta = \frac{\Delta I_C}{\Delta I_B} = \frac{11 \times 10^{-3}}{25 \times 10^{-6}} = 44$$

This simply means that a change in base current produces a change in collector current which is 44 times as large.

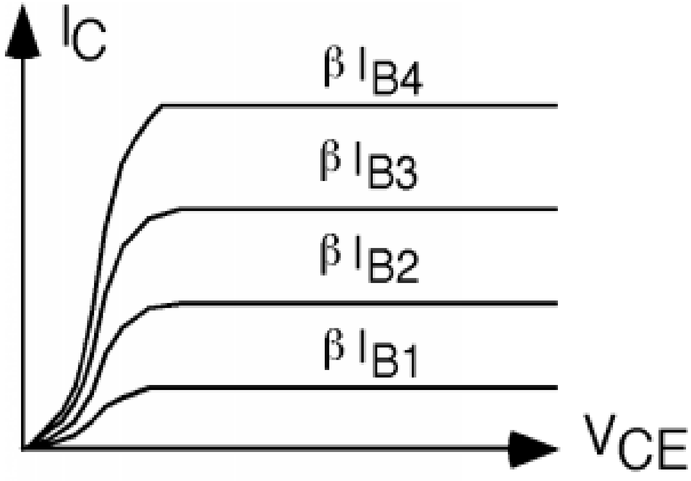
You may also see the term h_{fe} used in place of b. The terms h_{fe} and b are equivalent and may be used interchangeably. This is because "h_{fe}" means: h = hybrid (meaning mixture)

f = forward current transfer ratio
e = common emitter configuration

The resistance gain of the common emitter can be found in a method similar to the one used for finding beta:

$$R = \frac{R_{out}}{R_{in}}$$

Once the resistance gain is known, the voltage gain is easy to calculate since it is equal to the current gain (b) multiplied by the resistance gain (E = bR). And, the power gain is equal to the voltage gain multiplied by the current gain b (P = bE).



Common Collector

The common-collector configuration (CC) shown in figure 5.7.1 view C is used mostly for impedance matching. It is also used as a current driver, because of its substantial current gain. It is particularly useful in switching circuitry, since it has the ability to pass signals in either direction (bilateral operation).

In the common-collector circuit, the input signal is applied to the base, the output is taken from the emitter, and the collector is the element common to both input and output. The common collector is equivalent to our old friend the electron-tube cathode follower. Both have high input and low output resistance. The input resistance for the common collector ranges from 2 kilohms to 500 kilohms, and the output resistance varies from 50 ohms to 1500 ohms. The current gain is higher than that in the common emitter, but it has a lower power gain than either the common base or common emitter. Like the common base, the output signal from the common collector is in phase with the input signal. The common collector is also referred to as an emitter-follower because the output developed on the emitter follows the input signal applied to the base.

Transistor action in the common collector is similar to the operation explained for the common base, except that the current gain is not based on the emitter-to-collector current ratio, alpha (a). Instead, it is based on the emitter-to-base current ratio called GAMMA (g), because the output is taken off the emitter. Since a small change in base current controls a large change in emitter current, it is still possible to obtain high current gain in the common collector. However, since the emitter current gain is offset by the low output resistance, the voltage gain is always less than 1 (unity), exactly as in the electron-tube cathode follower

The common-collector current gain, gamma (g), is defined as

$$\gamma = \frac{\Delta I_E}{\Delta I_B}$$

and is related to collector-to-base current gain, beta (b), of the common-emitter circuit by the formula:

$$\gamma = \beta + 1$$

Since a given transistor may be connected in any of three basic configurations, there is a definite relationship, as pointed out earlier, between alpha (a), beta (b), and gamma (g). These relationships are listed again for your convenience:

$$\alpha = \frac{\beta}{\beta + 1} \quad \beta = \frac{\alpha}{1 - \alpha} \quad \gamma = \beta + 1$$

Take, for example, a transistor that is listed on a manufacturer's data sheet as having an alpha of 0.90. We wish to use it in a common emitter configuration. This means we must find beta. The calculations are:

$$\beta = \frac{\alpha}{1 - \alpha} = \frac{0.90}{1 - 0.90} = \frac{0.90}{0.1} = 9$$

Therefore, a change in base current in this transistor will produce a change in collector current that will be 9 times as large.

If we wish to use this same transistor in a common collector, we can find gamma (g) by:

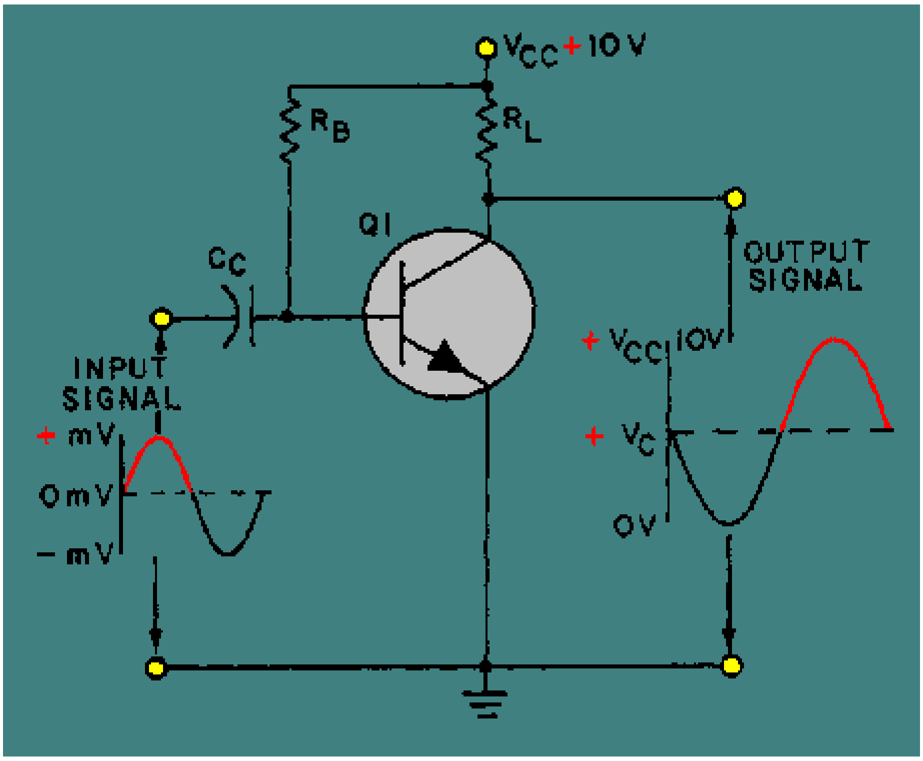
$$\gamma = \beta + 1 = 9 + 1 = 10$$

To summarize the properties of the three transistor configurations, a comparison chart is provided in table 6.7.1 for your convenience.

Table 5.7.1. - Transistor Configuration Comparison Chart

AMPLIFIER TYPE	**COMMON BASE**	**COMMON EMITTER**	**COMMON COLLECTOR**
INPUT/OUTPUT PHASE RELATIONSHIP	0°	180°	0°
VOLTAGE GAIN	HIGH	MEDIUM	LOW
CURRENT GAIN	LOW(a)	MEDIUM(b)	HIGH(g)
POWER GAIN	LOW	HIGH	MEDIUM
INPUT RESISTANCE	LOW	MEDIUM	HIGH
OUTPUT RESISTANCE	HIGH	MEDIUM	LOW

A reproduction of figure 5.6.1 is shown below for your convenience.



Transistor Configuration Comparison Chart

AMPLIFIER TYPE	COMMON BASE	COMMON EMITTER	COMMON EMITTER (Emitter Resistor)	COMMON COLLECTOR (Emitter Follower)
INPUT/OUTPUT PHASE RELATIONSHIP	**0°**	**180°**	**180°**	**0°**
VOLTAGE GAIN	**HIGH** $\frac{\alpha R_C}{R_s + r_e}$	**MEDIUM** $\frac{\beta(R_C \| r_o)}{R_s + r_\pi}$	**MEDIUM** $\frac{\beta R_C}{R_s + (\beta+1)(r_e + R_E)}$	**LOW** $\frac{(\beta+1)(R_L \| r_o)}{R_s + (\beta+1)[r_e + (R_L \| r_o)]}$
CURRENT GAIN	**LOW** α	**MEDIUM** $\beta \frac{r_o}{R_C + r_o}$	**MEDIUM** β	**HIGH** $(\beta+1)\frac{r_o}{r_o + R_L}$
POWER GAIN	**LOW**	**HIGH**	**HIGH**	**MEDIUM**
INPUT RESISTANCE	**LOW** r_e	**MEDIUM** $r_\pi = (\beta+1) r_e$	**MEDIUM** $(\beta+1)(r_e + R_E)$	**HIGH** $(\beta+1)[r_e + (r_o \| R_L)]$
OUTPUT RESISTANCE	**HIGH** R_C	**MEDIUM** $R_C \| r_o$	**MEDIUM** R_C	**LOW** $r_o \| \left[r_e + \frac{R_s}{(\beta+1)}\right]$

Input/Output Characteristics and AC Behavior

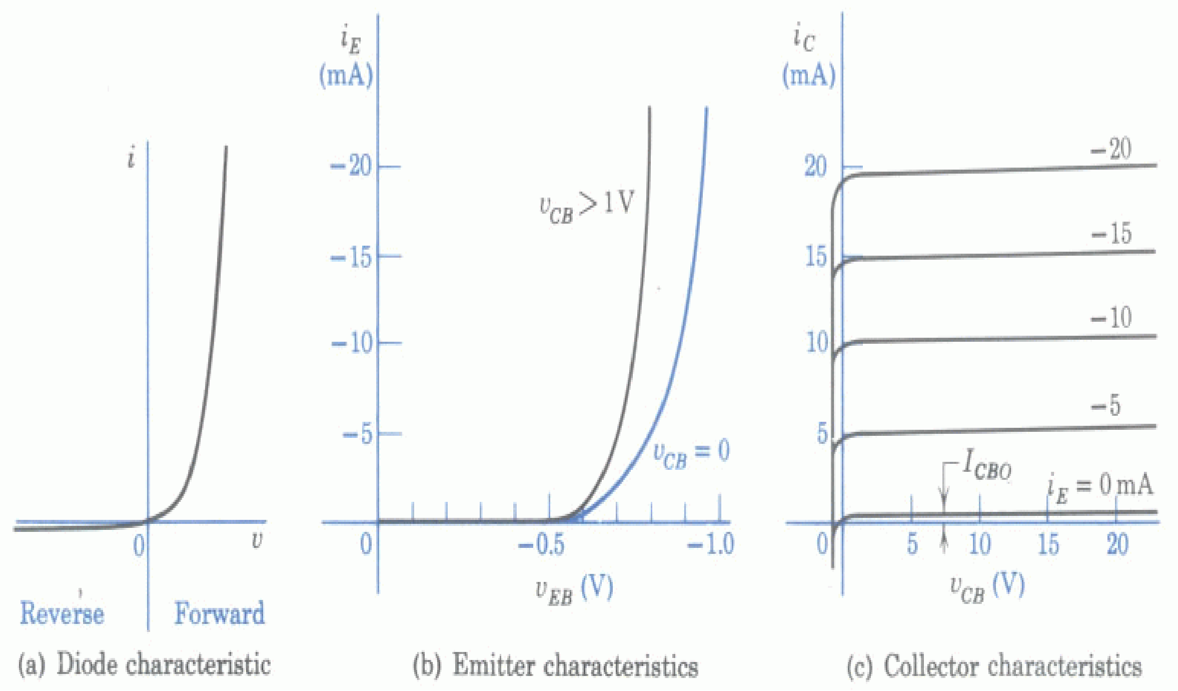
- **Common-Base:**
  - **Input characteristics:**

The EB junction is essentially the same as a forward biased diode, therefore the current-voltage characteristics is essentially the same as that of a diode:

$$I_E = I_0(e^{V_{BE}/V_T} - 1)$$

while the collector-base voltage $V_{CB} > 0$ also help enhance the current I_E to some extent.

  - **Output characteristics:**



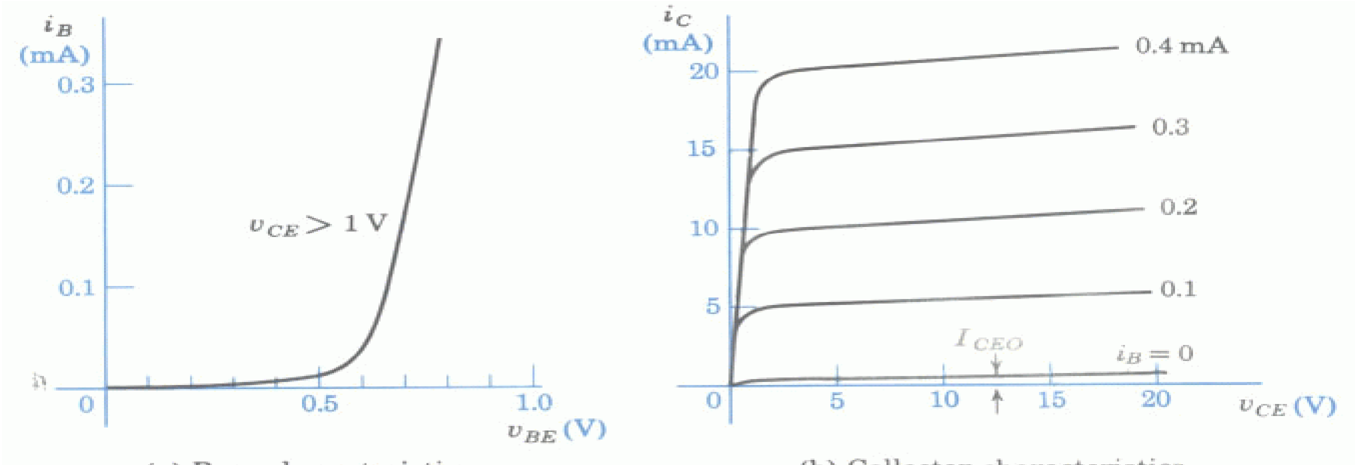
(a) Diode characteristic (b) Emitter characteristics (c) Collector characteristics

- **Common-Emitter:**
  - **Input characteristics:**

Same as in the case of common-base, the EB junction of common-emitter is also as a forward biased diode, the current-voltage characteristics is similar to that of a diode:

$$I_B = I_0(e^{V_{BE}/V_T} - 1)$$

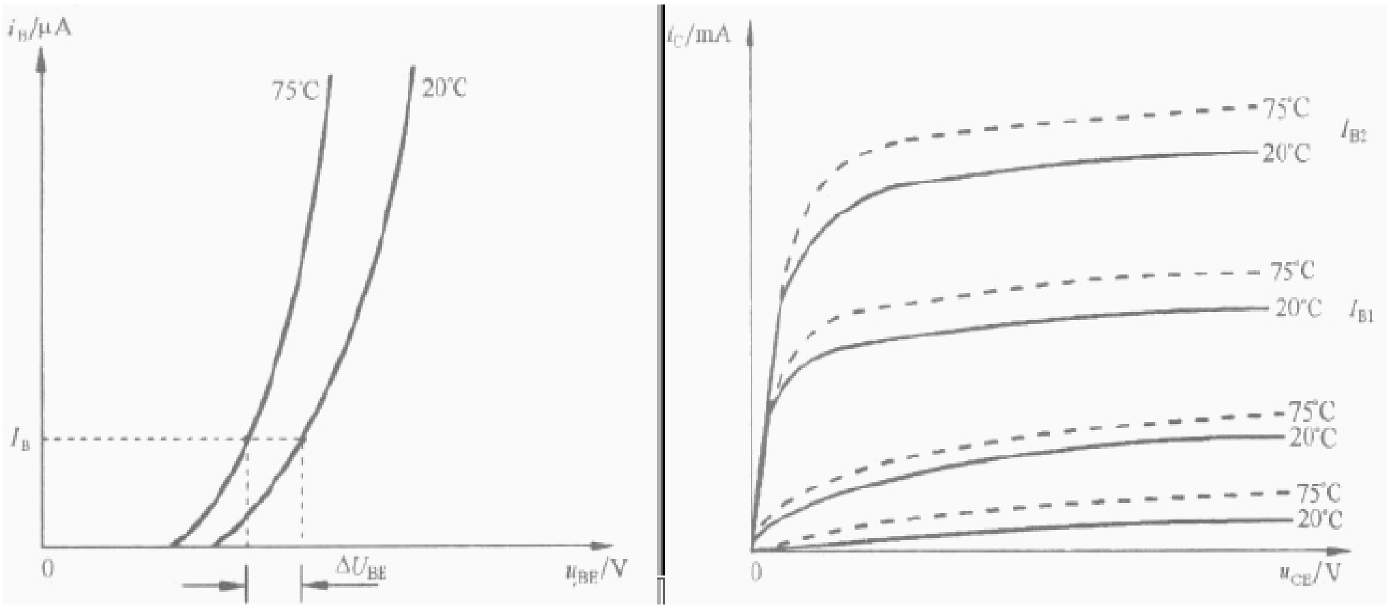
Output characteristics:



(a) Base characteristics (b) Collector characteristics

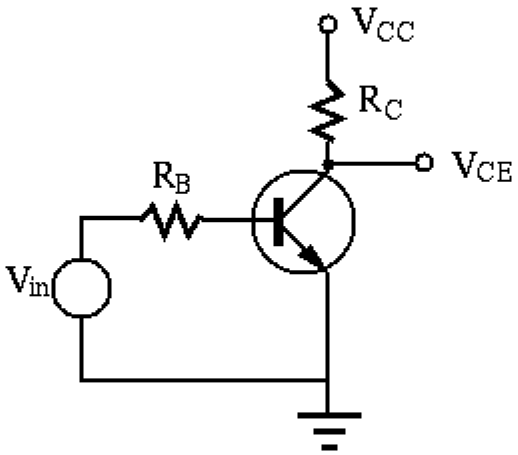
-
- Note that the common-base (CB) and common-emmitter (CE) configurations have similar collector characteristics with several differences:

○ In



Various parameters of a transistor change as functions of temperature. For example, β increases along with temperature.

Example: Assume in the CE circuit shown above, $V_{CC} = 15V$, $V_1 = 1V$, $R_B = 3K\Omega$, $R_C = 1.5K\Omega$, $\alpha = 0.9756$. Find output voltage V_2.



$$\beta = \alpha/(1-\alpha) = 40$$

- Find

I_B $\qquad$ $V_{BE} = 0.7\ V$

- Find . As the BE junction is forward biased, the voltage drop is about , and $I_B = (V_1 - V_{BE})/R_B = (1-0.7)/3 = 0.1\ mA$

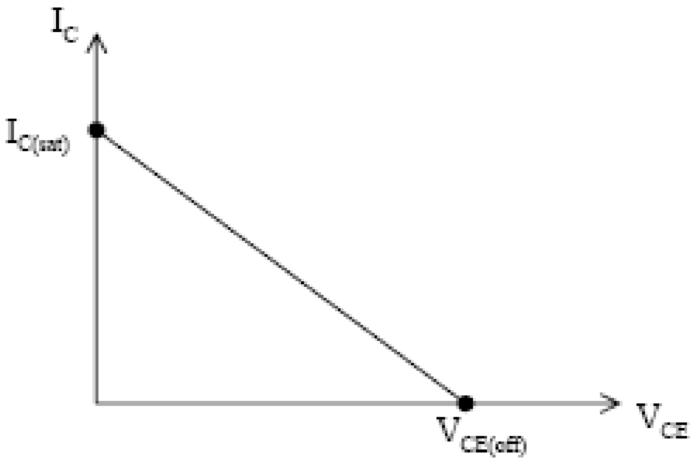
$$I_C = \beta I_B = 40 \times 0.1\ mA = 4\ mA$$

- Find

$$V_2 = V_{CC} - I_C R_C = 15\ V - 4\ mA \times 1.5\ K\Omega = 9\ V$$

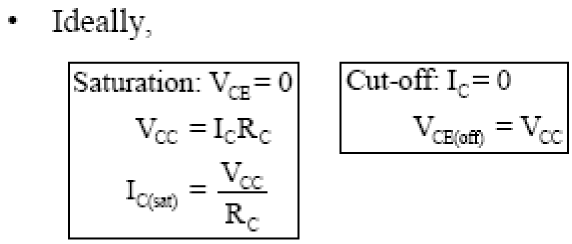
- Find
- **DC Load Line:**

Dc load line is a graph that represents the all possible combinations of I_C and V_{CE} for a given transistor biasing circuit. Which is given in fig



A generic dc load line

Fig 5.9.1



- Ideally,

Saturation: $V_{CE} = 0$

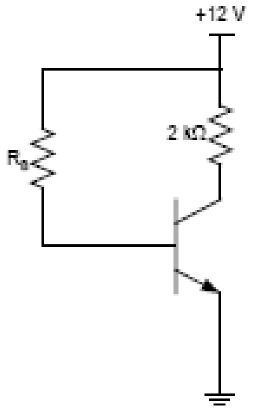
$$V_{CC} = I_C R_C$$

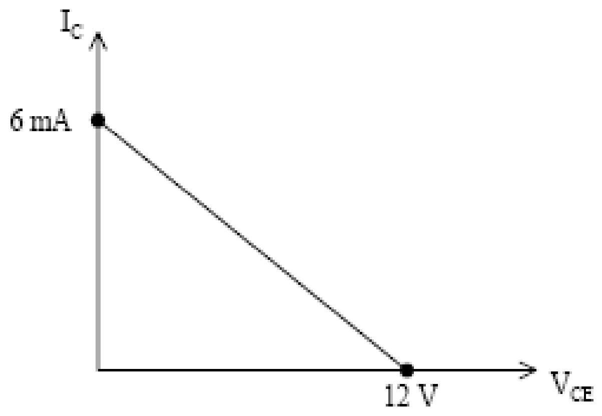
$$I_{C(sat)} = \frac{V_{CC}}{R_C}$$

Cut-off: $I_C = 0$

$$V_{CE(off)} = V_{CC}$$

Example 1:
Plot the dc load line for the circuit shown below.



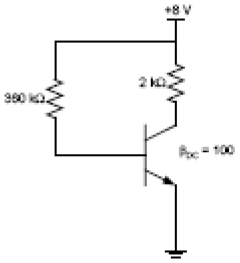


The dc load line

$$I_{C(sat)} = \frac{V_{CC}}{R_C} = \frac{12}{2k} = 6\text{ mA}$$

$$V_{CE(off)} = V_{CC} = 12\text{ V}$$

Q-Point:



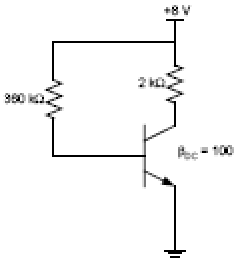
- When a BJT does not have an ac input, it will have specific dc values of I_C and V_{CE}.
- These values will correspond to a specific point on the dc load line.
- This point is called the Q-point. The letter Q comes from the word *quiescent*, meaning at rest.

- I_C and V_{CE} at Q-point are called I_{CQ} and V_{CEQ} respectively.
- Determine I_C and V_{CE} if V_{BE} is assumed to be 0.7 volt.

$$I_C = \frac{\beta_{DC}(V_{CC}-V_{BE})}{R_B}$$
$$= \frac{(100)(8-0.7)}{360k} = 2.028\text{ mA}$$

$$V_{CE} = V_{CC} - I_C R_C$$
$$= 8 - (2.028\text{mA})(2k) = 3.94\text{ V}$$

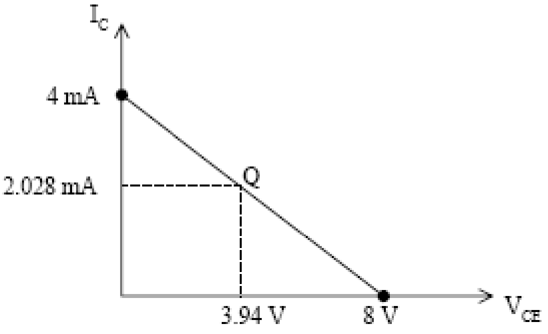
- Determine $I_{C(sat)}$ and $V_{CE(off)}$.



$$I_{C(sat)} = \frac{V_{CC}}{R_C} = \frac{8}{2k} = 4\text{ mA}$$

$$V_{CE(off)} = V_{CC} = 8\text{ V}$$

- Construct the dc load line and plot the Q-point.



The dc load line

- The circuit is said to be midpoint biased since the values of I_C and V_{CE} at Q-point are one-half of their maximum values.

Stability Factor(s):

The extent to which the collector current I_C is stabilized with varying I_{CO} is measured by a stability factor S. I defined as the rate of change of collector current I_C with respect to the collector base leakage current I_{CO}, kee both the current I_B and the current gain β constant

$$S = \frac{\partial I_C}{\partial I_{CO}} \sim \frac{dI_C}{dI_{CO}} \sim \frac{\Delta IC}{\Delta I_{CO}}, \ \beta \text{ and } I_B \text{ constant.} \qquad (5.11.1)$$

The collector current for a CE amplifier is given by

$$I_C = \beta I_B + (\beta+1) I_{CO} \qquad (5.11.2)$$

Differentiating the above equation with respect to I_C, we get

$$I = \beta \frac{dI_B}{dI_C} + (\beta+1)\frac{dI_{CO}}{dI_C} \qquad (5.11.3)$$

Therefore, $(1-\beta\frac{dI_B}{dI_C}) = \frac{(\beta+1)}{S}$

$$S = \frac{1+\beta}{1-\beta(\frac{dI_B}{dI_C})} \qquad (5.11.4)$$

$$S^{'} = \frac{\partial I_C}{\partial V_{BE}} \approx \frac{\Delta I_C}{\Delta V_{BE}} \qquad (5.11.5)$$

$$S^{''} = \frac{\partial I_C}{\partial \beta} \approx \frac{\Delta I_C}{\Delta \beta} \qquad (5.11.6)$$

Transistor Biasing

DC Biasing

The DC operating point of a transistor circuit need to be set up for it to work properly. The operating point is determined by the biasing circuit:

Fixed Bias Or Base Bias:

As shown in the circuit fig6.1a, two dc voltage supplies are needed to bias a BJT which is not practical.

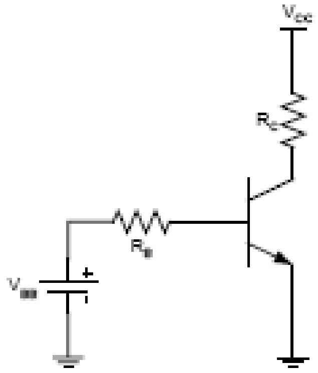


Fig 6.1a

In a simple biasing circuit 6.1b, V_{BB} eliminated by connecting the resistor R_B to the supply V_{CC}. This biasing circuit is called base bias, or fixed bias.

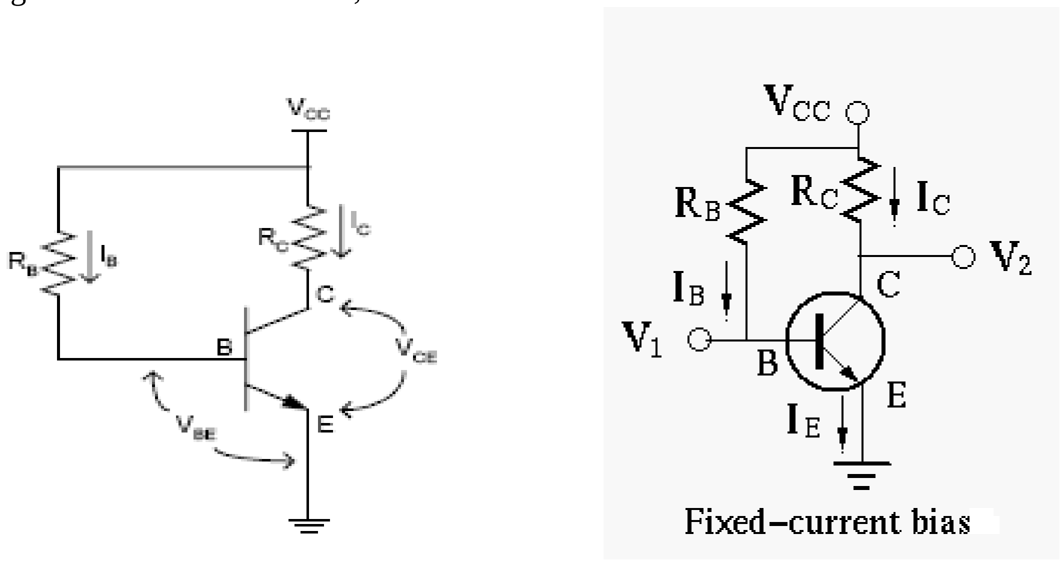


Fig 6.1.b

- As the voltage V_{BE} (0.7V) is small compared to V_{CC} (>10V), the base current can be estimated to be:

$$I_B = \frac{V_{CC} - V_{BE}}{R_B} \approx \frac{V_{CC}}{R_B} \tag{6.1}$$

The collector current is

$$I_C = \beta I_B + (\beta+1) I_{CB0} = \beta I_B + I_{CE0} \approx \beta \frac{V_{CC}-V_{BE}}{R_B} \approx \beta \frac{V_{CC}}{R_B} \qquad (6.2)$$

which is directly proportional to β. The output voltage is

$$V_{CE} = V_{CC} - I_C R_C \qquad (6.3)$$

As I_C and V_{CE} depend on β, which is different for different transistors and changes as a function of temperature, the operating point is unstable and inconsistent.

Stability factor:

Here $I_B = \frac{V_{CC}-V_{BE}}{R_B} \approx \frac{V_{CC}}{R_B}$, this equation is independent of current I_C, $dI_B/dI_C=0$.

So, S=(β+1)

Since β is large quantity, this is a very poor bias stable circuit.

Example 1: In the circuit of fixed current biasing, $V_{CC}=15V$, $R_B=2M\Omega$, $R_C=12K\Omega$. Assume $V_{BE}=0.7\ V$ Find the operating points (I_C, V_{CE}) for $\beta = 20,\ 100,\ 200$.

- When $\beta = 20$,

$$I_C = \beta I_B = \beta \frac{V_{CC}-V_{BE}}{R_B} = 20 \frac{15-0.7}{2\times 10^6} = 0.143mA$$

$$V_C = V_{CC} - I_C R_C = 15 - 0.143 \times 12 = 13.284V$$

When $\beta = 100$, $I_C = \beta I_B = \beta \frac{V_{CC}-V_{BE}}{R_B} = 100 \frac{15-0.7}{2\times 10^6} = 0.715mA$

$$V_C = V_{CC} - I_C R_C = 15 - 0.715 \times 12 = 6.42V$$

When $\beta = 200$, $I_C = \beta I_B = \beta \frac{V_{CC}-V_{BE}}{R_B} = 200 \frac{15-0.7}{2\times 10^6} = 1.43mA$

$$V_C = V_{CC} - I_C R_C = 15 - 1.43 \times 12 = -2.16V$$

How can there be a negative voltage while the voltage supply is $15V$? The collector current can no longer be determined by $I_C = \beta I_B$, as the maximum I_C corresponding to the transistor fully saturated with $V_{CE} = 0.2$ is $(V_{CC} - V_{CE})/R_C = 15.8/12 = 1.23\,mA$.

Collector-Feedback Bias:

A common emitter amplifier using collector- to –base bias circuit is shown in fig It provides some degree of stabilization to the amplifier operating point.
If the collector current I_C tends to increase due to either increase in temperature or the transistor has been replaced by the one with a higher β, the voltage drop across R_C increases, thereby reducing the value of $V_{CE.}$ Therefore, I_B decreases which, in turn, compensates the increase in I_C. Thus greater stability is obtained.

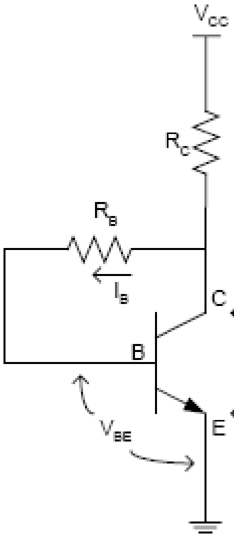


Fig6.2

- Applying KVL,

$$\begin{aligned} V_{CC} &= V_{RB} + V_{BE} + V_{RE} \\ &= I_B R_B + V_{BE} + (\beta_{DC}+1) I_B R_E \\ I_B &= \frac{V_{CC} - V_{BE}}{R_B + (\beta_{DC}+1) R_E} \\ I_{CQ} &= \frac{\beta_{DC}(V_{CC} - V_{BE})}{R_B + (\beta_{DC}+1) R_E} \\ V_{CEQ} &= V_{CC} - I_{CQ} R_C + I_E R_E \\ &\approx V_{CC} - I_{CQ}(R_C + R_E) \;;\; \beta_{DC} >> 1 \end{aligned}$$

Stability Factor:

$$V_{CC}=(I_B+I_C)R_C+I_B R_B+V_{BE} \qquad (6.2.1)$$

$$\frac{dI_B}{dI_C} = -\frac{R_E}{R_E + R_B} \quad I_B = \frac{V_{CC} - V_{BE} - I_C R_C}{R_C + R_B}$$

(6.2.2)

therefore , $\frac{dI_B}{dI_C} = -\frac{R_C}{R_C + R_B}$ (6.2.3)

Substituting equa.(6.2.3) in equa(5.11.4), we get

$$S = \frac{1+\beta}{1+\beta(\frac{R_C}{R_C + R_B})}$$

(6.2.4)

This value of the stability factor is smaller than the value obtained by fixed bias circuit. Also, S can be made small and the stability can be improved by making R_B small or R_C large.

Emitter-Feedback Bias/ Emitter Stabilized Bias Circuit:

The DC bias network of fig 6.3 contains an emitter resistor to improve the stability level over that of the fixed bias configuration.

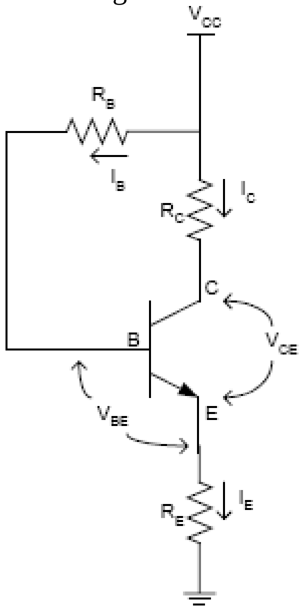


Fig 6.3

- Applying KVL,

$$\begin{aligned}
V_{CC} &= V_{RB}+V_{BE}+V_{RE} \\
&= I_B R_B + V_{BE} + (\beta_{DC}+1) I_B R_E \\
I_B &= \frac{V_{CC}-V_{BE}}{R_B+(\beta_{DC}+1)R_E} \\
I_{CQ} &= \frac{\beta_{DC}(V_{CC}-V_{BE})}{R_B+(\beta_{DC}+1)R_E} \\
V_{CEQ} &= V_{CC} - I_{CQ}R_C + I_E R_E \\
&\approx V_{CC} - I_{CQ}(R_C+R_E) \;;\; \beta_{DC} >> 1
\end{aligned}$$

As I_C and V_{CE} depend on β, which is different for different transistors and changes as a function of temperature, the operating point is unstable and inconsistent.

Stability Factor:

$$S = (\beta+1)\frac{1+{R_B}/{R_E}}{(\beta+1)+{R_B}/{R_E}}$$

When $\frac{R_B}{R_C} >> (\beta+1)$, then

$$S=\beta+1$$

When $\frac{R_B}{R_C} << (\beta+1)$, then

$$S=1$$

6.4 Voltage divider bias Circuit or Self Bias:

To correct the problem above, self-biasing circuit is used to decrease the effect of changing β by negative feed back.

Qualitatively, if I_C is increased due to increased β or temperature, the following happens:

$$I_C \uparrow \Longrightarrow V_E \uparrow \Longrightarrow V_{BE} \downarrow \Longrightarrow I_B \downarrow \Longrightarrow I_C = \beta I_B \downarrow$$

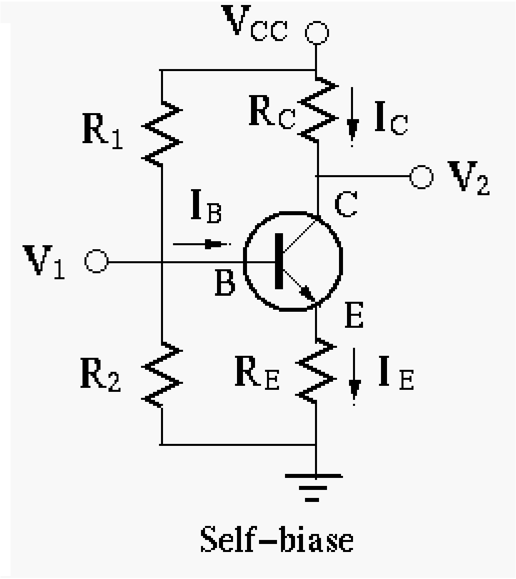


Fig 6.4

This is a negative feedback loop which tends to stabilize the operating point. To analyze this circuit quantitativelyWith the help of fig6.5, we first find the base voltage V_B and base current I_B. Note that only when the base current is much smaller than the current through R_2 ($I_B \ll I_2$).

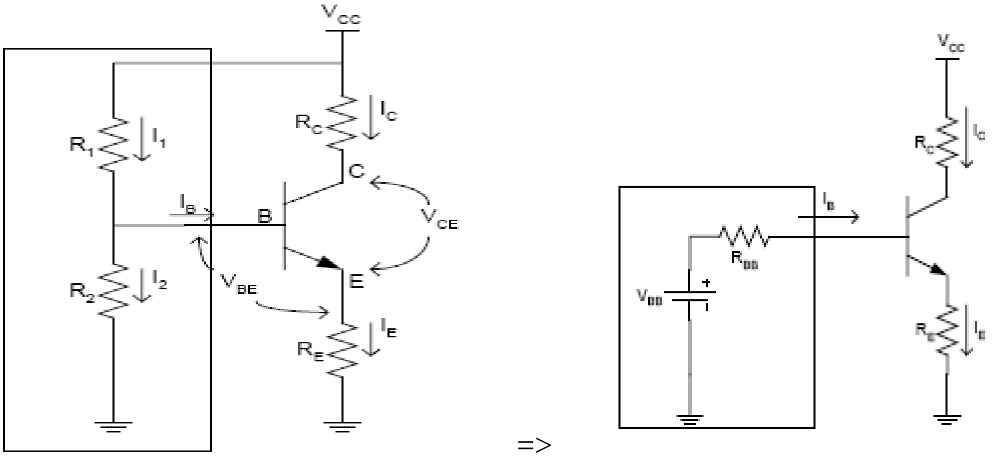


Fig 6.5

Then can we approximate V_B by voltage divider as:

$$V_B = V_{CC} \frac{R_2}{R_1 + R_2}$$

If the condition $I_B \ll I_2$ is not satisfied, we have to use Thevenin's theorem to replace the base circuit by a voltage source

$$V_{BB} = \frac{R_2}{R_1 + R_2} V_{CC}$$

in series with a resistance

$$R_B = \frac{R_1 R_2}{R_1 + R_2}$$

Where R_B is a Thevenin resistance calculate by fig 6.6.

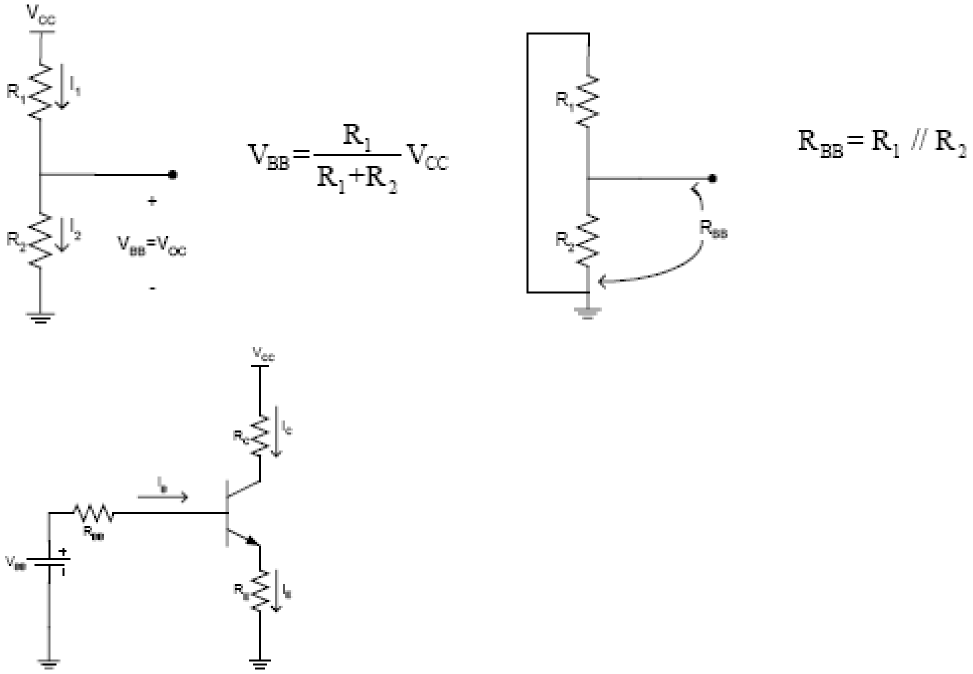


Fig 6.6

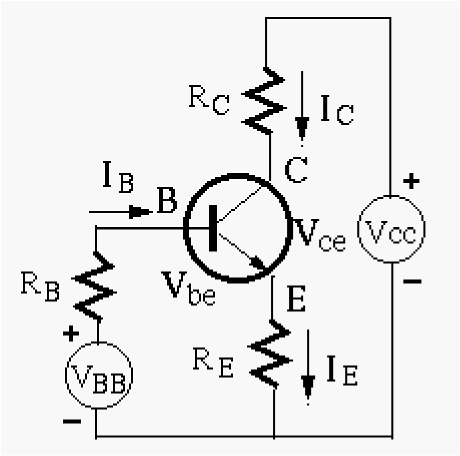


Fig 6.7

Next we use KVL to the base loop at the fig 6.7, to get

$$V_{BB} - I_B R_B - V_{BE} - (I_C + I_B) R_E = V_{BB} - V_{BE} - I_C R_E - I_B (R_B + R_E) = 0$$

Substituting $I_C = \beta I_B$

we get $V_{BB} - V_{BE} = [(\beta + 1)R_E + R_B]I_B$

which can be solved to obtain both I_B and I_C:

$$I_B = \frac{V_{BB} - V_{BE}}{(\beta + 1)R_E + R_B}, \qquad I_C = \beta I_B = \frac{\beta(V_{BB} - V_{BE})}{(\beta + 1)R_E + R_B}$$

If we assume that even for the minimum possible β, it is still true that $R_B \ll \beta_{min}R_E$, e.g., $R_B = 0.1 \times \beta_{min}R_E$, then I_C can be approximated as $I_C \approx \frac{V_{BB} - V_{BE}}{R_E}$ i.e., I_C, and thereby V_{CE} and the DC operating point is determined only by the resistors of the circuit, independent of the β value which may change for different transistors or at different temperatures. Comparing this with fixed biasing with $I_C \approx \beta(V_{CC} - V_{BE})/R_B$ directly proportional to β, the self-biasing circuit has a much more stable operating point.

Stability Factor:

$$V_{BB} - V_{BE} = [(\beta + 1)R_E + R_B]I_B$$

$$\frac{dI_B}{dI_C} = -\frac{R_E}{R_E + R_B}$$

$$S = \frac{1+\beta}{1+\beta(\frac{R_E}{R_E + R_B})}$$

Or, $S = (\beta + 1)\frac{1 + {}^{R_B}\!/_{R_E}}{(\beta + 1) + {}^{R_B}\!/_{R_E}}$

When $\frac{R_B}{R_C} >> (\beta + 1)$, then

S=β+1

When $\frac{R_B}{R_C} << (\beta + 1)$, then

S=1

Example 2: In the circuit of self-biasing, $V_{CC} = 28V$, $R_1 = 90K\Omega$, $R_2 = 10K\Omega$, $R_E = 2K\Omega$, $R_C = 14K\Omega$, Assume $V_{BE} = 0.7$. Find the operating points for $\beta = 20, 100, 200$.

$$R_B = R_1 R_2/(R_1 + R_2) = 10 \times 90/(10 + 90) = 9K\Omega$$

$$V_{BB} = V_{CC} R_2/(R_1 + R_2) = 28 \times 10/(10 + 90) = 2.8V$$

When $\beta = 20$, $I_C = \beta I_B = \beta \frac{V_{BB} - V_{BE}}{(\beta + 1)R_E + R_B} = \frac{20 \times (2.8 - 0.7)}{21 \times 2000 + 9000} \approx 0.82mA$

$$V_C = V_{CC} - I_C R_C = 28 - 0.82 \times 10^{-3} \times 14 \times 10^3 = 16.5V$$

$$V_E = I_E R_E = (I_C + I_E)R_E = (\beta + 1) I_B R_E \approx 0.82mA2K\Omega = 1.64V$$

$$V_{CE} = V_C - V_E = 16.5 - 1.64 = 14.8V$$

When $\beta = 100$,

$$I_C = \beta I_B = \beta \frac{V_{BB} - V_{BE}}{(\beta + 1)R_E + R_B} = \frac{100 \times (2.8 - 0.7)}{101 \times 2000 + 9000} = 2.1/2110 \approx 1mA$$

$$V_C = V_{CC} - I_C R_C = 28 - 10^{-3} \times 14 \times 10^3 = 14V$$

$$V_E = I_E R_E = (I_C + I_E)R_E = (\beta + 1) I_B R_E \approx 1mA2K\Omega = 2V$$

$$V_{CE} = V_C - V_E = 12V$$

When $\beta = 200$, $I_C = \beta I_B = \beta \frac{V_{BB} - V_{BE}}{(\beta + 1)R_E + R_B} = \frac{200 \times (2.8 - 0.7)}{201 \times 2000 + 9000} \approx 1.02mA$

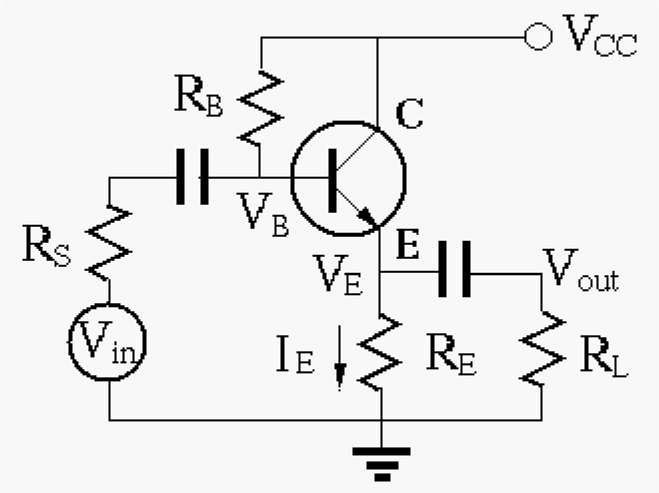
$$V_C = V_{CC} - I_C R_C = 28 - 1.02 \times 10^{-3} \times 14 \times 10^3 = 13.7V$$

$$V_E = I_E R_E = (I_C + I_E)R_E = (\beta + 1) I_B R_E \approx 1.02mA2K\Omega = 2V$$

$$V_{CE} = V_C - V_E = 13.7 - 2 = 11.7V$$

Emitter Follower

An emitter follower circuit shown in the figure is widely used in AC amplification circuits. The input and output of the emitter follower are the base and the emitter, respectively, therefore this circuit is also called common-collector circuit.



Emitter follower circuit

DC operating point

$$\begin{cases} V_{CC} = R_B I_B + V_{be} + R_E I_E \\ I_E = (\beta+1) I_B \\ V_{CE} = V_{CC} - R_E I_E \end{cases}$$

Solving these equations, we can get I_B, I_E and V_{CE}

$$I_B = \frac{V_{CC} - V_{BE}}{R_B + (\beta+1)R_E}$$

$$I_C = \beta I_B = \frac{\beta(V_{CC} - V_{BE})}{R_B + (\beta+1)R_E} \approx \frac{V_{CC} - V_{BE}}{R_E} \qquad \text{if } R_B \ll (\beta+1)R_E$$

$$V_{CE} = V_{CC} - R_E I_E$$

SMALL AND LARGE SIGNAL MODELS OF TRANSISTOR

BJT circuit models

A large variety of bipolar junction transistor models have been developed. One distinguishes between small signal and large signal models. We will discuss here first the hybrid pi model, a small signal model, which lends itself well to small signal design and analysis. The next model is the charge control model, which is particularly well suited to analyze the large-signal transient behavior of a bipolar transistor. And we conclude with the derivation of the SPICE model parameters.

Small signal model (hybrid pi model)

The hybrid pi model of a BJT is a small signal model, named after the “π”-like equivalent circuit for a bipolar junction transistor. The model is shown in Figure 7.1.1. It consists of an input impedance, r_π, an output impedance r_0, and a voltage controlled current source described by the transconductance, g_m. In addition it contains the base-emitter capacitances, the junction capacitance, $C_{j,BE}$, and the diffusion capacitance, $C_{d,BE}$, and the base-collector junction capacitance, $C_{j,BC}$, also referred to as the Miller capacitance.

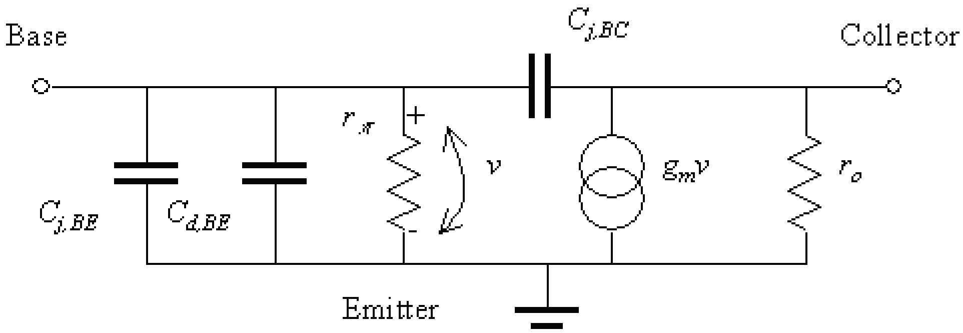


Figure 7.1.1. : Small signal model (hybrid pi model) of a bipolar junction transistor.

The transconductance, g_m, of a bipolar transistor is defined as the change in the collector current divided by the change of the base-emitter voltage.

$$g_m \overset{\Delta}{=} \frac{\partial I_C}{\partial V_{BE}} = \frac{I_C}{nV_t} \tag{7.1.1}$$

The base input resistance, r_π, is defined as the change of the emitter-base voltage divided by the change of the base current.

$$r_\pi \overset{\Delta}{=} \frac{\partial V_{BE}}{\partial I_B} = \beta \frac{\partial V_{BE}}{\partial I_C} = \frac{\beta}{g_m} = \frac{nV_t}{I_B} \tag{7.1.2}$$

The output resistance, r_o, is defined as:

$$r_o \overset{\Delta}{=} \frac{\partial V_{CE}}{\partial I_C} \cong \frac{\partial V_{CB}}{\partial I_C} = \frac{|V_A|}{I_C} \tag{7.1.3}$$

The base-emitter and base-collector junction capacitances are given by:

$$C_{j,BE} = \frac{C_{j,BE0}}{\sqrt{1 - \frac{V_{BE}}{\phi_{i,BE}}}} \qquad (7.1.4)$$

$$C_{j,BC} = \frac{C_{j,BC0}}{\sqrt{1 - \frac{V_{BC}}{\phi_{i,BC}}}} \qquad (7.1.5)$$

for the case where the base-emitter and base-collector junctions are abrupt. Since the base-emitter is strongly forward biased in the forward active mode of operation, one has to also include the diffusion capacitance of the base:

$$C_{d,BE} = \frac{I_E}{V_t}\tau_B \qquad (7.1.6)$$

Based on the small signal model shown in Figure 7.1.1, we can now calculate the small signal current gain versus frequency, h_{fe}, of a BJT biased in the forward active mode and connected in a common emitter configuration. The maximum current gain is calculated while shorting the output, resulting in:

$$h_{fe} = \frac{i_C}{i_B} = \frac{g_m}{i_B}v = \frac{\beta}{1 + j\omega(C_{j,BE} + C_{d,BE})r_\pi} \qquad (7.1.7)$$

The unity gain frequency, f_T, also called the transit frequency is obtained by setting the small signal current gain, h_{fe}, equals to one, resulting in:

$$|\frac{i_C}{i_B}| = 1 \cong \frac{\beta}{2\pi\ f_T\,(C_{j,BE} + C_{d,BE})r_\pi} \qquad (7.1.8)$$

This transit frequency can be expressed as a function of the transit time, τ:

$$f_T = \frac{1}{2\pi\ \tau} \qquad (7.1.9)$$

Where the transit time, τ, equals:

$$\tau = \frac{C_{j,BE}\ nV_t}{I_E} + \frac{w_B'^{\,2}}{2D_{n,B}} = \tau_E + \tau_B \qquad (7.1.10)$$

The circuit model therefore includes the charging time of the base-emitter capacitance, τ_E, as well as the base transit time, τ_B, but not the transit time of the carriers through the base-collector depletion region, τ_C.

$$\tau_C = \frac{t_C}{2} = \frac{x_{d,BC}}{2v_{sat}} \qquad (7.1.11)$$

The total transit time then becomes:

$$\tau = \frac{C_{j,BE}\; nV_t}{I_E} + \frac{w_B^{'\,2}}{2D_{n,B}} + \frac{x_{d,BC}}{2v_{sat}} = \tau_E + \tau_B + \tau_C \qquad (7.1.12)$$

While the unity gain frequency, f_T, is an important figure of merit of a bipolar transistor, another even more important figure of merit is the maximum oscillation frequency, f_{MAX}. This figure of merit predicts the unity power gain frequency and as a result indicates the maximum frequency at which useful power gain can be expected from a device. The maximum oscillation frequency, f_{MAX}, is linked to the transit frequency, f_T, and is obtained from:

$$f_{\max} = \sqrt{\frac{f_T}{2\pi\; R_B C_{j,BC}}} \qquad (7.1.13)$$

Where R_B is the total base resistance and $C_{j,BC}$ is the base-collector capacitance. The total base resistance consists of the series connection of metal-semiconductor contact resistance, the resistance between the base contact metal and the emitter and the intrinsic base resistance. Assuming a base contact, which is longer than the penetration depth this base resistance equals:

$$R_B = \frac{\sqrt{R_{s,c}\,\rho_c}}{L_{s,E}} + R_{s,BE}\;\frac{\Delta L}{L_{s,E}} + \frac{1}{3}\,R_s\,\frac{W_{s,E}}{L_{s,E}} \qquad (7.1.14)$$

for a one-sided base contact, where $R_{s,c}$, $R_{s,BE}$ and R_s are the sheet resistances under the base contact, between the base contact and the emitter and underneath the emitter respectively. $L_{s,E}$ is the emitter stripe length of the emitter, $W_{s,E}$ is the emitter stripe width of the emitter and Δ_L is the alignment distance between the base contact and emitter. For a double-sided base contact, the total base resistance equals

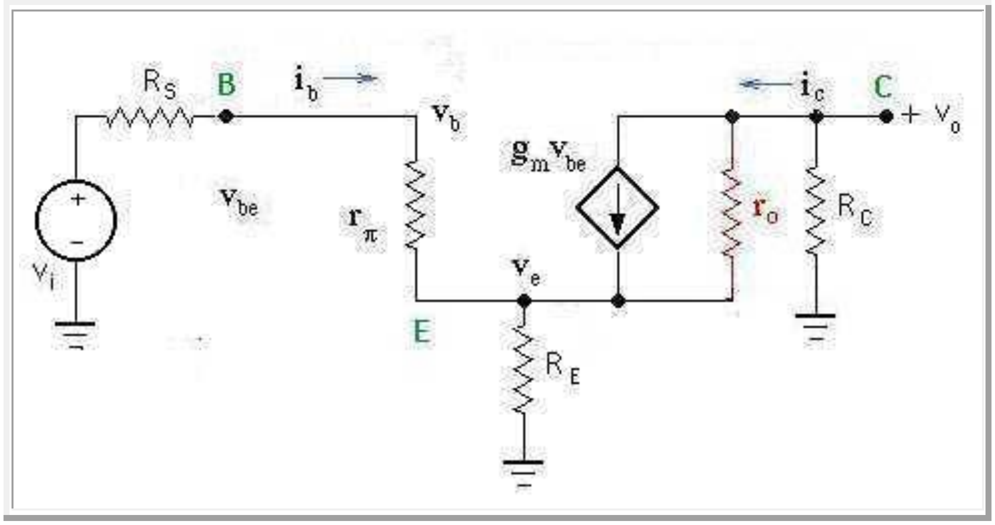
$$R_B = \frac{\sqrt{R_{s,c}\,\rho_c}}{2L_{s,E}} + R_{s,BE}\;\frac{\Delta L}{2L_{s,E}} + \frac{1}{12}\,R_s\,\frac{W_{s,E}}{L_{s,E}} \qquad (7.1.15)$$

The base-collector capacitance equals:

$$C_{j,BC} = \varepsilon_s\,\frac{A_C}{x_{d,BC}} \qquad (7.1.16)$$

Where A_C is the base-collector area.

CE Amplifier Small-Signal Equivalent Circuit:



To analyze this configuration, we first set down the complete nodal equations:

For B node: $$\mathbf{v}_b \frac{1}{\mathrm{R}_s \,\|\, \mathrm{r}_\pi} = \mathbf{v}_i \frac{1}{\mathrm{R}_s} + \mathbf{v}_e \frac{1}{\mathrm{r}_\pi}$$

For C node: $$\mathbf{v}_e \frac{1}{\mathrm{r}_o} = \mathbf{v}_o \frac{1}{\mathrm{R}_C \,\|\, \mathrm{r}_o} + g_m\left(\mathbf{v}_b - \mathbf{v}_e\right)$$

For E node: $$\mathbf{v}_e \frac{1}{\mathrm{R}_E \,\|\, \mathrm{r}_o \,\|\, \mathrm{r}_\pi} = \mathbf{v}_o \frac{1}{\mathrm{r}_o} + \mathbf{v}_b \frac{1}{\mathrm{r}_\pi} + + g_m\left(\mathbf{v}_b - \mathbf{v}_e\right)$$

Using the relationship $g_m \, \mathrm{r}_\pi = \beta$, the nodal equations can be rewrite in a more homogeneous form:

For B node: $$\mathbf{v}_b \frac{1}{\mathrm{R}_s \,\|\, \mathrm{r}_\pi} = \mathbf{v}_i \frac{1}{\mathrm{R}_s} + \mathbf{v}_e \frac{1}{\mathrm{r}_\pi}$$

For C node: $$\mathbf{v}_e \left\{\frac{1}{\mathrm{r}_o} + \frac{\beta}{\mathrm{r}_\pi}\right\} = \mathbf{v}_o \frac{1}{\mathrm{R}_C \,\|\, \mathrm{r}_o} + \mathbf{v}_b \frac{\beta}{\mathrm{r}_\pi}$$

For E node: $$\mathbf{v}_e \left\{\frac{1}{\mathrm{R}_E \,\|\, \mathrm{r}_o \,\|\, \mathrm{r}_\pi} + \frac{\beta}{\mathrm{r}_\pi}\right\} = \mathbf{v}_o \frac{1}{\mathrm{r}_o} + \mathbf{v}_b \frac{\beta + 1}{\mathrm{r}_\pi}$$

Eliminating v_o from the last two nodal equations we find that

$$\mathbf{v}_e \left\{\frac{\mathrm{r}_o}{\mathrm{R}_E \,\|\, \mathrm{r}_o} - \frac{\mathrm{R}_C \,\|\, \mathrm{r}_o}{\mathrm{r}_o} + \frac{\mathrm{r}_o}{\mathrm{r}_\pi}\left[1 + \frac{\beta\, \mathrm{r}_o}{\left(\mathrm{R}_C + \mathrm{r}_o\right)}\right]\right\} = \mathbf{v}_b \frac{\mathrm{r}_o}{\mathrm{r}_\pi}\left[1 + \frac{\beta\, \mathrm{r}_o}{\left(\mathrm{R}_C + \mathrm{r}_o\right)}\right]$$

and if we substitute this expression into the first nodal equation we find that

$$\mathbf{v}_b = \mathbf{v}_i \, \frac{R_s \parallel r_\pi}{R_s} \left\{ \frac{r_o}{R_E \parallel r_o} - \frac{R_C \parallel r_o}{r_o} + \frac{r_o}{r_\pi} \left[1 + \frac{\beta \, r_o}{(R_C + r_o)} \right] \right\} \left\{ \frac{r_o}{R_E \parallel r_o} - \frac{R_C \parallel r_o}{r_o} + \frac{r_\pi}{(R_s + r_\pi)} \frac{r_o}{r_\pi} \left[1 + \frac{\beta \, r_o}{(R_C + r_o)} \right] \right\}$$

$$\mathbf{v}_e = \mathbf{v}_i \, \frac{r_o}{r_\pi} \frac{R_s \parallel r_\pi}{R_s} \left[1 + \frac{\beta \, r_o}{(R_C + r_o)} \right] \left\{ \frac{r_o}{R_E \parallel r_o} - \frac{R_C \parallel r_o}{r_o} + \frac{r_\pi}{(R_s + r_\pi)} \frac{r_o}{r_\pi} \left[1 + \frac{\beta \, r_o}{(R_C + r_o)} \right] \right\}^{-1}$$

Finally, substituting these two expressions into the second nodal equation we find the following expression for the voltage gain:

$$A_V \equiv \frac{\mathbf{v}_o}{\mathbf{v}_i} = -\frac{\beta \, R_C \parallel r_o}{(R_s + r_\pi)} \left\{ 1 - (\beta + 1) \frac{(\beta + 1)}{\beta} \frac{R_E \parallel r_o}{r_o} \right\} \left\{ 1 - \frac{R_E \parallel r_o}{r_o} \frac{R_C \parallel r_o}{r_o} + \frac{r_o}{r_\pi} \frac{R_E \parallel r_o}{r_o} \frac{R_s \parallel r_\pi}{R_s} \left[1 + \beta \frac{R_C \parallel r_o}{R_C} \right] \right.$$

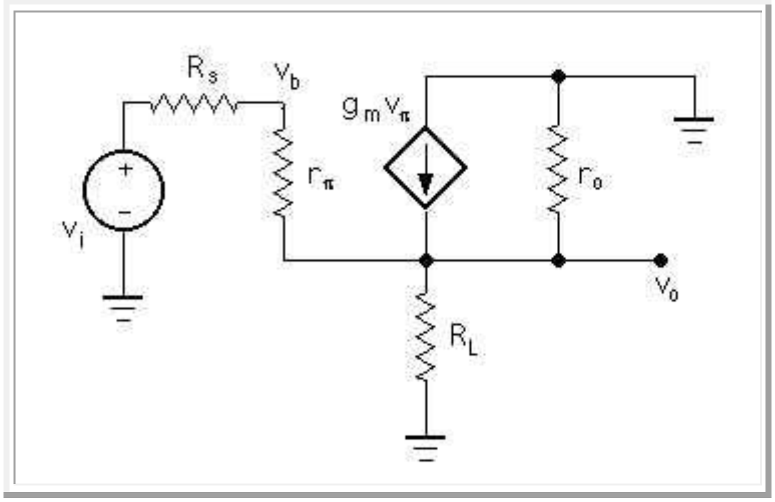
When $R_E = 0$ this expression reduces to

$$A_V \equiv \frac{\mathbf{v}_o}{\mathbf{v}_i} = -\beta \, \frac{R_C \parallel r_o}{R_s + r_\pi}$$

When $R_E \neq 0$ but $r_o \Rightarrow \infty$ it reduces to

$$A_V \equiv \frac{\mathbf{v}_o}{\mathbf{v}_i} = -\frac{\beta \, R_C}{R_s + (\beta + 1)(R_E + r_e)}$$

CE ("Emitter-Follower") Amplifier Small-Signal Equivalent Circuit



Again to analyze this configuration, we first set down the complete nodal equations:

For B node: $$\mathbf{v}_b \, \frac{1}{R_s \parallel r_\pi} = \mathbf{v}_i \, \frac{1}{R_s} + \mathbf{v}_o \, \frac{1}{r_\pi}$$

For E node: $$\mathbf{v}_o \, \frac{1}{R_L \parallel r_o \parallel r_\pi} = \mathbf{v}_b \, \frac{1}{r_\pi} + g_m \left(\mathbf{v}_b - \mathbf{v}_o \right)$$

Again using the relationship $g_m \, r_\pi = \beta$, the nodal equations can be rewrite in a more homogeneous form:

For B node: $$v_b \frac{1}{R_s \| r_\pi} = v_i \frac{1}{R_s} + v_o \frac{1}{r_\pi}$$

For E node: $$v_o \left[\frac{1}{R_L \| r_o \| r_\pi} + \frac{\beta}{r_\pi} \right] = v_b \frac{\beta + 1}{r_\pi}$$

Substituting the second nodal equation into the first we find the following expression for the voltage gain:

$$A_V \equiv \frac{v_o}{v_i} = \frac{(\beta+1)\, R_L \| r_o}{R_s + r_\pi + (\beta+1)\, R_L \| r_o} = \frac{(\beta+1)\, R_L \| r_o}{R_s + (\beta+1)\left(r_e + R_L \| r_o\right)}$$

A "trickly" calculation is required to obtain the output impedance. To do so we first shut off the input voltage and then apply test voltage source, v_x, to the output terminal. Under these circumstances, the current into the output terminal is given by:

$$i_x = \frac{v_x}{r_o \| (r_\pi + R_s)} - g_m\, v_\pi = \frac{v_x}{r_o \| (r_\pi + R_s)} - \frac{\beta}{r_\pi}\left[-\frac{r_\pi}{(r_\pi + R_s)} v_x \right]$$
$$= v_x \left[\frac{1}{r_o} + \frac{1}{(r_\pi + R_s)}(\beta+1) \right]$$

Therefore, the relatively low output impedance is given by:

$$\frac{1}{R_o} \equiv \frac{i_x}{v_x} = \frac{1}{r_o} + \frac{1}{(r_\pi + R_s)}(\beta+1)$$

or $$R_o = r_o \| \left(\frac{r_\pi + R_s}{\beta+1} \right) = r_o \| \left(r_e + \frac{R_s}{\beta+1} \right) \Rightarrow \left(\frac{r_\pi + R_s}{\beta+1} \right)$$

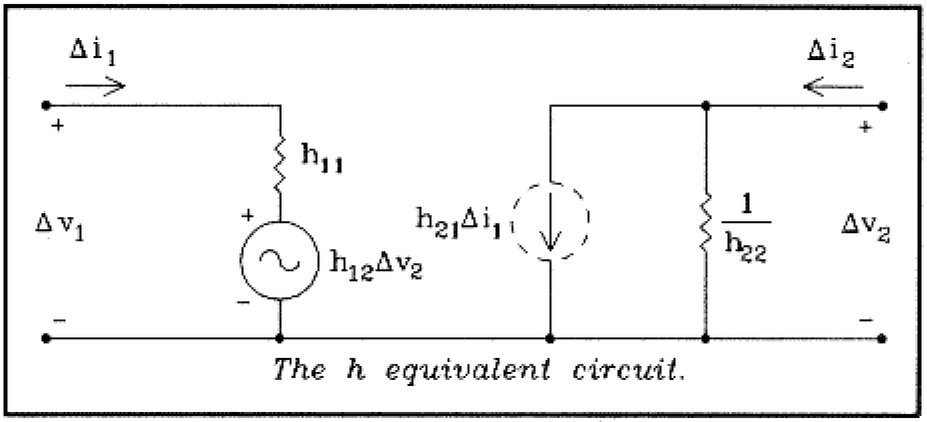
while the relatively high input impedance is given by:

$$\frac{1}{R_i} = \frac{i_i}{v_i} = \frac{i_i}{v_i} \frac{v_i - v_o}{R_s + r_\pi} = \frac{1}{R_s + r_\pi}\left[1 - \frac{v_o}{v_i}\right] = \frac{1}{R_s + r_\pi}\left[1 - \frac{(\beta+1)\, R_L \| r_o}{R_s + (\beta+1)\left(r_e + R_L \| r_o\right)}\right]$$

or $$R_i = R_s + (\beta+1)\left(r_e + R_L \| r_o\right) = R_s + r_\pi + (\beta+1)\, R_L \| r_o \quad \Rightarrow (\beta+1)\, R_L$$

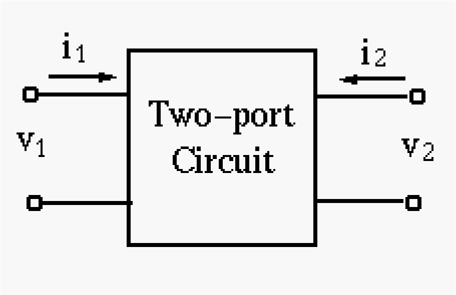
7.2. Small signal model (h parameter model):

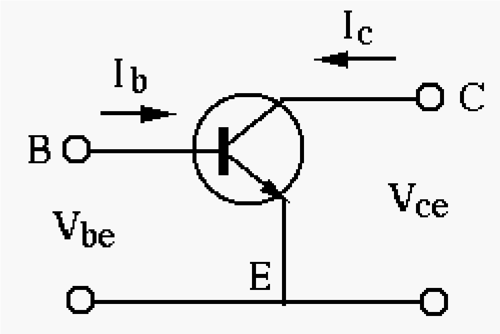
The simpler models have been developed to predict (at least approximately) how a bipolar transistor will perform as a small-signal amplifier. A model frequently used by semiconductor manufacturers is the *hybrid* model, and the values used with it are called *hybrid* or "*h*" parameters:



The h equivalent circuit.

Two-port circuit:





$$h_{11} = \left.\frac{\Delta v_1}{\Delta i_1}\right|_{\Delta v_2 = 0} \qquad h_{12} = \left.\frac{\Delta v_1}{\Delta v_2}\right|_{\Delta i_1 = 0}$$

$$h_{21} = \left.\frac{\Delta i_2}{\Delta i_1}\right|_{\Delta v_1 = 0} \qquad h_{22} = \left.\frac{\Delta i_2}{\Delta v_2}\right|_{\Delta i_1 = 0}$$

Each of these h-parameters is defined as the ratio of a particular *response* of the transistor divided by a certain *excitation* which causes that response .

A transistor circuit can be treated as a two-port circuit with input and output ports with four variables (v_1, i_1, v_2, i_2). In general two of the four variables are independent and the rest two can be expressed as their functions:

$$\begin{cases} v_1 = f_1(i_1, i_2) \\ v_2 = f_2(i_1, i_2) \end{cases} \quad \text{or} \quad \begin{cases} i_1 = f_3({}_1, v_2) \\ i_2 = f_4(v_1, v_2) \end{cases} \quad \text{or} \quad \begin{cases} v_1 = f_5(i_1, v_2) \\ i_2 = f_6(i_1, v_2) \end{cases}$$

We use the third *hybrid* model to describe the CE transistor circuit with $v_1 = v_{be}$, $i_1 = i_b$, $v_2 = v_{ce}$, and $i_2 = i_{ce}$:

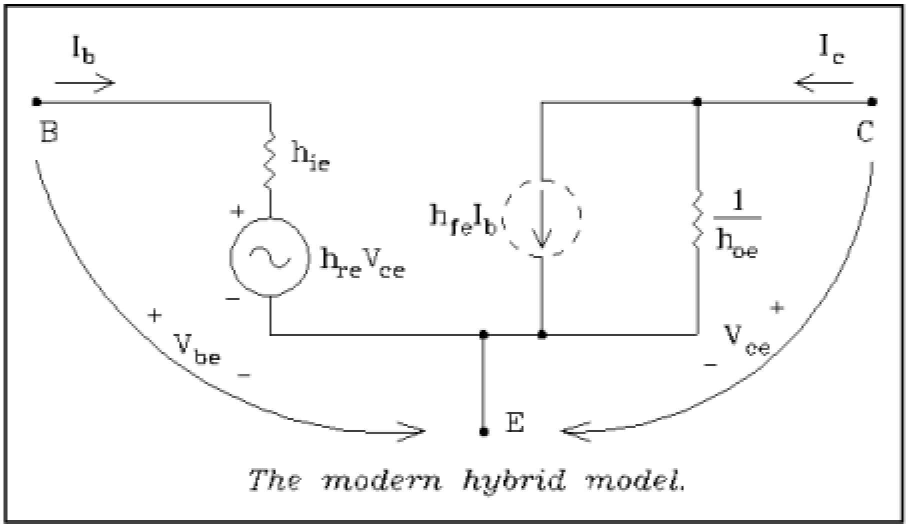
$$\begin{cases} v_{be} = v_{be}(i_b, v_{ce}) \\ i_c = i_c(i_b, v_{ce}) \end{cases}$$

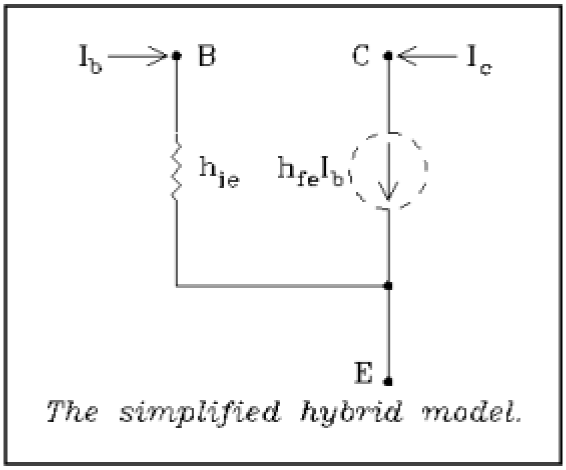
Taking the total derivative, we get:

$$dv_{be} = \frac{\partial v_{be}}{\partial i_b} di_b + \frac{\partial v_{be}}{\partial v_{ce}} dv_{ce} = h_i di_b + h_r dv_{ce} \qquad di_c = \frac{\partial i_c}{\partial i_b} di_b + \frac{\partial i_c}{\partial v_{ce}} dv_{ce} = h_f di_b + h_o dv_{ce}$$

$\mathbf{h_{11}} = \mathbf{h_{ie}}$: *dynamic input resistance* in the common emitter configuration.
$\mathbf{h_{12}} = \mathbf{h_{re}}$: *reverse feedback* from the output circuit - perhaps due to the bulk resistance of the emitter region.
$\mathbf{h_{21}} = \mathbf{h_{fe}}$: *forward current gain* of the transistor.
$\mathbf{h_{22}} = \mathbf{h_{oe}}$: *output conductance.*

7.2.1 Small Signal Analysis of Common Emitter configuration of a transistor(h-parameter model)





The relationship between the voltages and currents for a transistor in the common emitter configuration is shown in figure 7.2.1

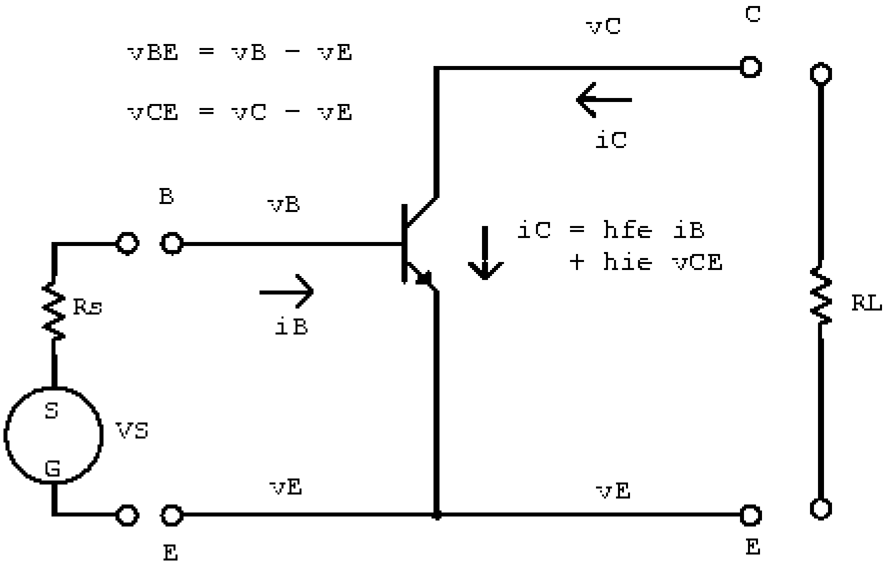
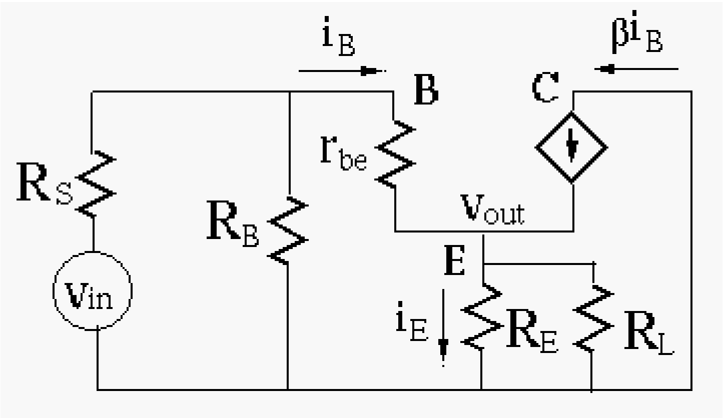


Figure 7.2.1: Transistor in the common emitter configuration.

AC small-signal equivalent circuit



Small AC signal model

Large signal model (Charge control model)

The charge control model of a bipolar transistor is an extension of the charge control model of a p-n diode. Assuming the “short” diode model to be valid, one can express the device currents as a function of the charges in each region, divided by the corresponding transit or lifetime. In the general case one considers the forward bias charges as well as the reverse bias charges. This results in:

$$I_E = \frac{\Delta Q_{p,E}}{t_{r,E}} + \frac{\Delta Q_{n,B,f}}{t_{r,B}} - \frac{\Delta Q_{n,B,r}}{t_{r,B}} + \frac{\Delta Q_{n,B,r}}{\tau_{n,B}} \tag{7.3.1}$$

$$I_B = \frac{\Delta Q_{p,E}}{t_{r,E}} + \frac{\Delta Q_{n,B,f}}{\tau_{n,B}} + \frac{\Delta Q_{p,C}}{t_{r,C}} + \frac{\Delta Q_{n,B,r}}{\tau_{n,B}} \tag{7.3.2}$$

$$I_C = -\frac{\Delta Q_{p,C}}{t_{r,C}} - \frac{\Delta Q_{n,B,r}}{t_{r,B}} + \frac{\Delta Q_{n,B,f}}{t_{r,B}} - \frac{\Delta Q_{n,B,f}}{\tau_{n,B}} \tag{7.3.3}$$

Under forward active mode of operation, this model can be simplified since the reverse mode components can be ignored. A transient model can be obtained by adding the rate of change of the charges over time. To further simplify the model, we also ignore the minority carrier charge, $\Delta Q_{p,E}$, in the emitter. This results in the following equations:

$$I_E = \frac{\Delta Q_{n,B,f}}{t_{r,B}} + \frac{dQ_{n,B,f}}{dt} \tag{7.3.4}$$

$$I_B = \frac{\Delta Q_{n,B,f}}{\tau_{n,B}} + \frac{1}{\beta_F + 1}\frac{dQ_{n,B,f}}{dt} \tag{7.3.5}$$

$$I_C = \frac{\beta_F}{\beta_F + 1}[\Delta Q_{n,B,f} + \frac{dQ_{n,B,f}}{dt}] \tag{7.3.6}$$

As an example we now apply this charge control model to the abrupt switching of a bipolar transistor. Consider the circuit shown in Figure 7.3.1.(a). As one applies a positive voltage to the base, the base-emitter junction will become forward biased so that the collector current will start to rise. The input is then connected to a negative supply voltage, V_R. This reverses the base current and the base-emitter junction capacitance is discharged. After this transient, the transistor is eventually turned off and the collector current reduces back to zero. A full analysis would require solving the charge control model equations simultaneously, while adding the external circuit equations. Such approach requires numeric simulation tools.

To simplify this analysis and provide insight, we now assume that the base current is constant before and after switching. This approximation is very good under forward bias since the base-emitter voltage is almost constant. Under reverse bias, the base current will vary as the base-emitter voltage varies, but conceivably one could design a circuit that does provide a constant reverse current.

The turn-on of the BJT consists of an initial delay time, $t_{d,1}$, during which the base-emitter junction capacitance is charged. This delay is followed by the increase of the collector current, quantified by the rise time, t_{rise}. This rise time is obtained by applying the charge control equation for the base current, while applying a base current I_{BB} with the voltage source V_{BB}:

$$I_{BB} = \frac{\Delta Q_{n,B,f}}{\tau_{n,B}} + \frac{1}{\beta_F + 1}\frac{d\Delta Q_{n,B,f}}{dt} \tag{7.3.7}$$

where

$$I_{BB} = \frac{V_{BB} - V_{BE}}{R_B} \cong \frac{V_{BB} - 0.7}{R_B} \quad \text{(for a silicon BJT)} \tag{7.3.8}$$

This differential equation can be solved resulting in:

$$\Delta Q_{n,Bf} = I_{BB}\tau_B[1 - \exp(\frac{t}{t_{r.B}})] \tag{7.3.9}$$

If the device does not reach saturation, the charge reaches its steady state value with a time constant $t_{r,B}$, which equals the base transit time of the BJT. The corresponding collector current will be proportional to the excess minority carrier charge until the device reaches saturation or:

$$\begin{aligned} I_C &= \frac{I_{BB}\tau_B}{t_{r,B}}[1 - \exp(\frac{t}{t_{r.B}})] \text{ for } \Delta Q_{n,Bf} \le \frac{V_{CC} - V_{sat}}{R_L} t_{r,B} \\ I_C &= \frac{V_{CC} - V_{sat}}{R_L} \text{ for } \Delta Q_{n,Bf} \ge \frac{V_{CC} - V_{sat}}{R_L} t_{r,B} \end{aligned} \tag{7.3.10}$$

A larger base voltage, V_{BB}, will therefore result in a larger charging current, I_{BB}, which in turn decreases the rise time and causes the BJT to saturate more quickly. There also will be more excess minority carrier charge stored in the base region after the BJT is turned on. The rise time, t_{rise}, is then obtained by finding the time when the saturation current is reached or:

$$t_{rise} = t_{r,B}\ln(1 - \frac{V_{CC} - V_{sat}}{R_L I_{BB}\tau_{n,B}} t_{r,B}) \tag{7.3.11}$$

While switching back to the negative power supply, V_R, the base current is reversed. As long as significant charge is still stored in the base region, the collector current will continue to exist. Only after this excess charge is removed, will the base-emitter junction capacitor be discharged and the BJT be turned off. The removal of the excess charge can take a significant delay time labeled as $t_{d,2}$ on the figure. Again we can calculate the time evolution of the excess charge and calculate the collector current from it. To first order the delay time, $t_{d,2}$, equals:

$$t_{d2} = \frac{I_{BB}\,\tau_{n,B}}{I_R} \quad \text{with } I_R \cong \frac{V_R}{R_B} \tag{7.3.12}$$

This delay time can be significantly larger than the rise time t_{rise}. Also note that a higher base turn-on current I_{BB} results in a larger turn-off delay as more minority carrier charge is stored in the base.

The actual fall time, t_f, depends on the remaining storage charge at the onset of saturation as well as the charge stored by the base-emitter junction capacitance.

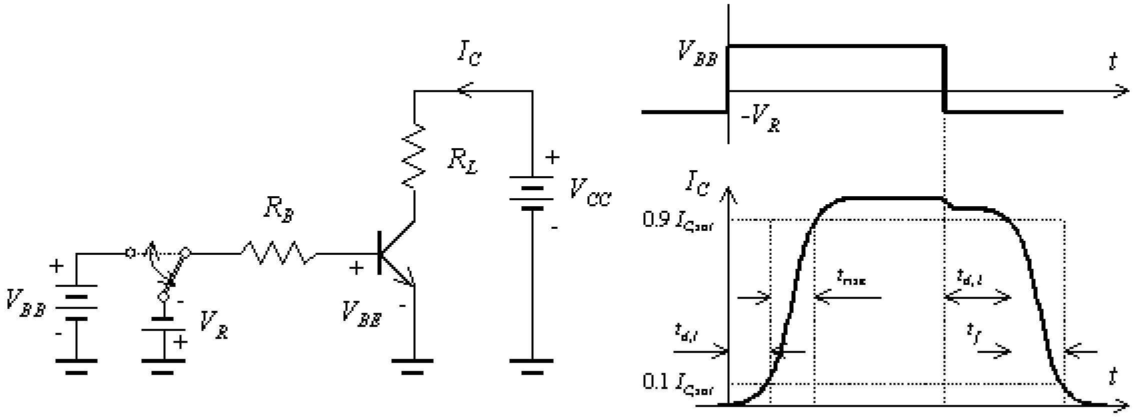


Figure 7.3.1. : Switching behavior of a BJT: a) bias circuit used to explain the switching behavior. b) Applied voltage and resulting collector current.

Module IV: Junction Field Effect Transistor (JFET)

The single channel junction field-effect transistor (JFET) is the simplest transistor. As shown in the schematics below (Figure 1) for the n-channel JFET (left) and the p-channel JFET (right), these devices are simply an area of doped silicon with two diffusions of the opposite doping.

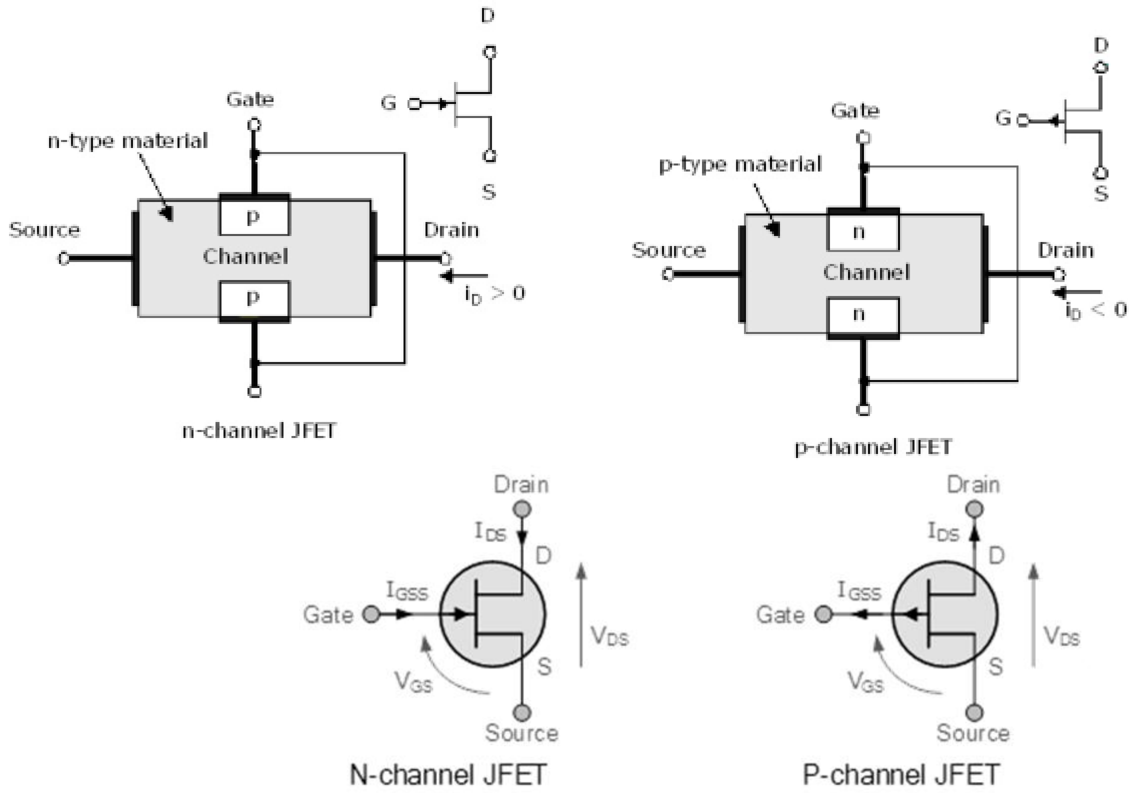


Fig 1

Like the BJT, the JFET is a three terminal device. Although there are physically two gate diffusions, they are tied together and act as a single gate terminal. The other two contacts, the drain and source, are placed at either end of the channel region. The JFET is a symmetric device (the source and drain may be interchanged).

The operation of the JFET is based on controlling the bias on the pn junction between gate and channel. If a voltage is applied between the drain and source, current will flow (the conventional direction for current flow is from the terminal designated to be the gate to that which is designated as the source). The device is therefore in a normally on state. To turn it off, we must apply an appropriate voltage to the gate and use the depletion region created at the junction to control the channel width.

The following discussion is going to focus on the n-channel JFET. He operating and characteristics of the p-channel JFET may then be deduced by making the necessary changes to voltage polarities and current directions.

- The source of the JFET will provide a common ground for all device terminals.
- The voltage applied to the drain will be designated V_{DD}
- The voltage applied to the gate will be designated V_{GG}

The n-channel JFET connected to the voltages sources V_{DD} is presented to the right (Figure 2(a)). The polarity indicated for V_{DD} means that the electrons in the channel will be attracted to the drain and conventional current will flow in the direction indicated by I_D (negative charges moving in a negative direction have the same effect as positive charges moving in a positive direction). The polarity voltage applied to the gate (V_{GG}) ensures that the gate channel pn junction will be reverse biased and essentially no current will flow (the reverse bias saturation current is considered negligible). The JFET is a voltage-controlled device, with two controlling voltages (V_{DD} and V_{GG}).

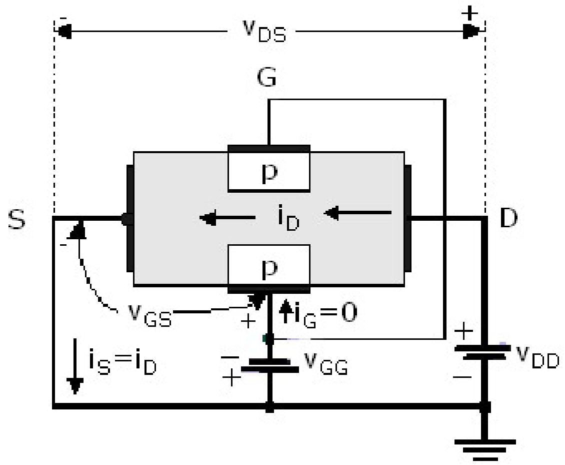
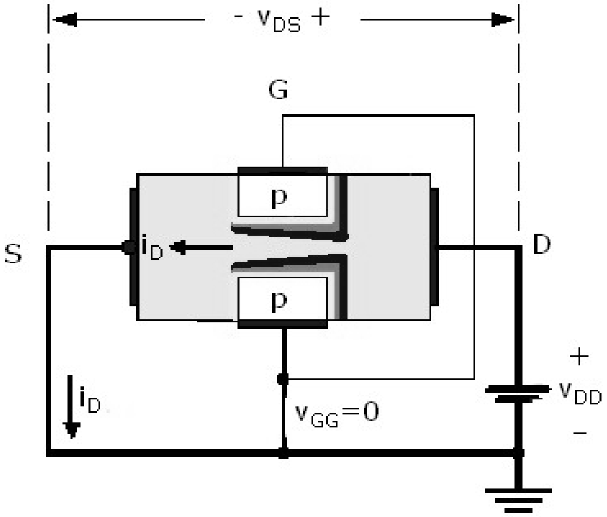


Fig 2(a)

JFET Drain-To-Source Voltage Variation



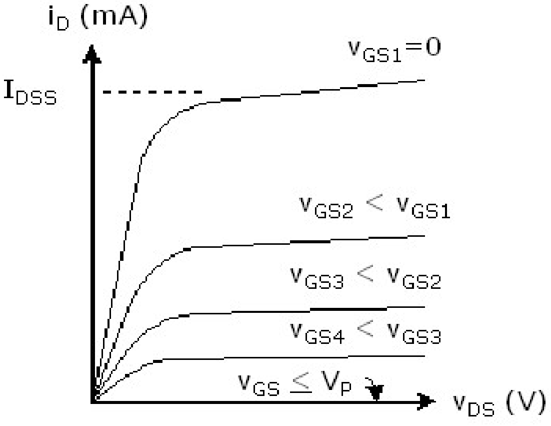
For a constant V_{GG}, the effect of v variation may be illustrated as follows:

- V_{DD} must be greater than zero for current to flow in the direction defined in an n-channel JFET. The applied V_{DD}, or the voltage between drain and source (ground) appears as a voltage drop across the length of the channel, with the voltage increasing along the channel from source to drain (i.e., V_{DS} =0 at the source, V_{DS}= V_{DD} at the drain). We're also going to assume a constant doping so that the voltage variation in the channel is linear.
- When V_{DD} is very small, the voltage variation in the channel is very small and it has no effect on the channel shape. For this case, the depletion region is only due to the pn junction as shown in figure.
- As V_{DD} increases, the increasing potential at the drain reverse biases the pn junctions. Since the voltage drop across the channel increases from source to drain, the reverse bias of the pn junction also increases from source to drain. Since the depletion region is a function of bias, the depletion region also gets wider from source to drain, causing the channel to become tapered as shown in red in the figure. The current still increases with increasing V_{DD} , however there is no longer a linear relationship between V_{DD} and I_{DD} since the channel resistance is a function of its width.
- Further increases in V_{DD} result in a more tapered shape to the channel and increasing nonlinearities in the I_D-V_{DD} relationship.
- This process continues until a V_{DS} is reached where the depletion regions from the pn junctions merge. Analytically, this occurs when the gate-to drain voltage V_{GD} is less than some threshold V_T and is known as the pinch-off point. At this point, the drain current saturates and further increases in V_{DS} little (ideally zero) change in i_D. For the case V_{GG}= V_{GS} = 0, the drain current at pinch-off is called the drain-source saturation current, $I_{DSS.}$ Operation beyond the pinch-off point (V_{DS} > V_{GS} - V_T for an n-channel device) defines the normal operating or saturation region of the JFET.

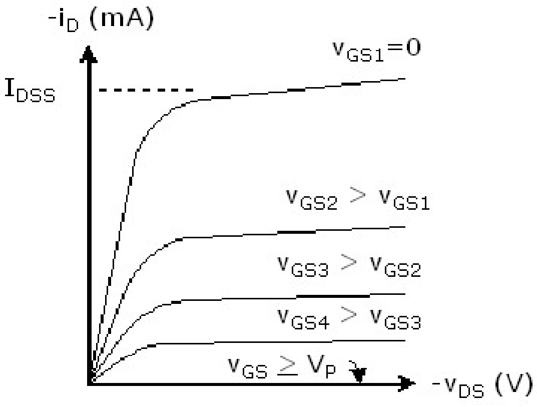
JFET Gate-To-Source Voltage Variation

For a constant V_{GG}, varying V_{DS} yields a single i_D-V_{DS} characteristic curve. To develop a characteristic curves for the JFET device, we need to look at the effect of V_{GS} variation. The figure shows a simple illustration of the variation of V_{GG} with a constant (and small)

$V_{DD.}$ If V_{DD} is small, V_{DS} is small and the channel width is essentially constant. In the figure, the increasing width of the pn junction depletion region is due to the increasing reverse bias of the junction resulting from the application of a negative V_{GG} of increasing magnitude. As the depletion widths increase, the channel width decreases, resulting in a lower conductivity (higher resistivity) of the channel. As V_{GS} is made more negative (for an n-channel device), a value of V_{GS} is reached for which the channel is completely depleted (no free carriers) and no current will flow regardless of the applied V_{DS}. This is called the threshold, or pinch-off, voltage and occurs at V_{GS}= $V_{GS\ (OFF)}$. The threshold voltage for an n-channel JFET is negative ($V_T < 0$).



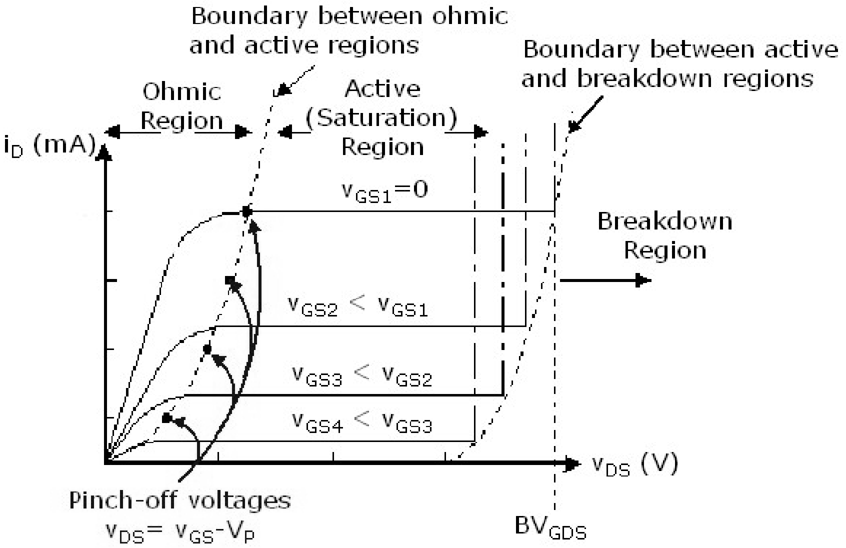
V-I characteristics n-channel JFET.



V-I characteristics p-channel JFET.

JFET Transfer Characteristics

The transfer characteristic for a FET device is presented as a plot curves representing drain current as a function of drain-to-source voltage for a sequence of constant gate-to-source voltages, as illustrated in the figures above for the n-channel and p-channel JFET. After the JFET reaches saturation, i_D remains relatively constant with a very small slope for further increases in V_{DS} (the slope of the curves would be zero for an ideal device).



Transfer characteristic for an n-channel device

Below pinch-off, the channel essentially behaves like a constant resistance. This linear region of operation is called ohmic (or sometimes triode), and is where the JFET may be used as a resistor whose value is determined by the value of $V_{GS.}$ The transistor may function as a variable resistor when operated in the ohmic region by varying the value of V_{GS}. However, the magnitude of V_{GS} increases, the range of V_{DS} where the transistor may be operated as an ohmic resistor decreases. In the ohmic region, the potentials at all three terminals strongly affect the drain current, and the drain current obeys the relationship:

$$i_D = K\left[2(v_{GS} - V_T)v_{DS} - v_{DS}^2\right], \text{ where } K = \frac{I_{DSS}}{V_P^2}$$

The drain current in the saturation region may be defined by using the
Shockley equation in terms of the drain-source saturation current (I_{DSS}), the threshold voltage (V_T) and the applied gate-to-source voltage (V_{GS}) as:

$$i_D \cong I_{DSS}\left(1 - \frac{v_{GS}}{V_T}\right)^2$$

As V_{DS} continues to increase, a point is reached when the drain-to-source voltage becomes so large that avalanche breakdown occurs at the drain end of the gate-channel junction. At the breakdown points, shown by dashed lines in the figure, i_D increases sharply with negligible increases in $V_{DS.}$ The value of V_{DS} denoted BV_{GDS} is the breakdown voltage of the pn junction (i.e., when (V_{GS} =0). As can be seen in the figure above, the breakdown voltage is also a function of V_{GS} – as the magnitude of V_{GS} increases (more negative for n-channel and more positive for p-channel) the breakdown voltage decreases.

Metal Oxide Field Effect Transistor (MOSFET):

As well as the Junction Field Effect Transistor (JFET), there is another type of Field Effect Transistor available whose Gate input is electrically insulated from the main current carrying channel and is therefore called an **Insulated Gate Field Effect Transistor** or IGFET. The most common type of insulated gate FET which is used in many different types of electronic circuits is called the **Metal Oxide Semiconductor Field Effect Transistor** or **MOSFET** for short.

The **IGFET** or **MOSFET** is a voltage controlled field effect transistor that differs from a JFET in that it has a "Metal Oxide" Gate electrode which is electrically insulated from the main semiconductor n-channel or p-channel by a very thin layer of insulating material usually silicon dioxide, commonly known as glass.

This ultra thin insulated metal gate electrode can be thought of as one plate of a capacitor. The isolation of the controlling Gate makes the input resistance of the **MOSFET** extremely high way up in the Mega-ohms (MΩ) region thereby making it almost infinite.

As the Gate terminal is isolated from the main current carrying channel "NO current flows into the gate" and just like the JFET, the MOSFET also acts like a voltage controlled resistor were the current flowing through the main channel between the Drain and Source is proportional to the input voltage. Also like the JFET, the MOSFETs very high input resistance can easily accumulate large amounts of static charge resulting in the **MOSFET** becoming easily damaged unless carefully handled or protected.

Types of MOSFET:

The MOSFET is classified into two types such as;

- Depletion mode MOSFET
  - (a) N-channel
  - (b) P-channel
- Enhancement mode MOSFET
  - (a) N-channel
  - (b) P-channel

Enhancement-mode MOSFET:

The more common Enhancement-mode MOSFET or eMOSFET, is the reverse of the depletion-mode type. Here the conducting channel is lightly doped or even undoped making it non-conductive. This results in the device being normally "OFF" (non-conducting) when the gate bias voltage, VGS is equal to zero. The circuit symbol shown above for an enhancement MOS transistor uses a broken channel line to signify a normally open non-conducting channel.
For the n-channel enhancement MOS transistor a drain current will only flow when a gate voltage (VGS) is applied to the gate terminal greater than the threshold voltage (VTH) level in which conductance takes place making it a transconductance device.

The application of a positive (+ve) gate voltage to a n-type eMOSFET attracts more electrons towards the oxide layer around the gate thereby increasing or enhancing (hence its name) the thickness of the channel allowing more current to flow. This is why this kind of transistor is called an enhancement mode device as the application of a gate voltage enhances the channel.
Increasing this positive gate voltage will cause the channel resistance to decrease further causing an increase in the drain current, ID through the channel. In other words, for an n-channel enhancement mode MOSFET: +VGS turns the transistor “ON”, while a zero or -VGS turns the transistor “OFF”. Then, the enhancement-mode MOSFET is equivalent to a “normally-open” switch.

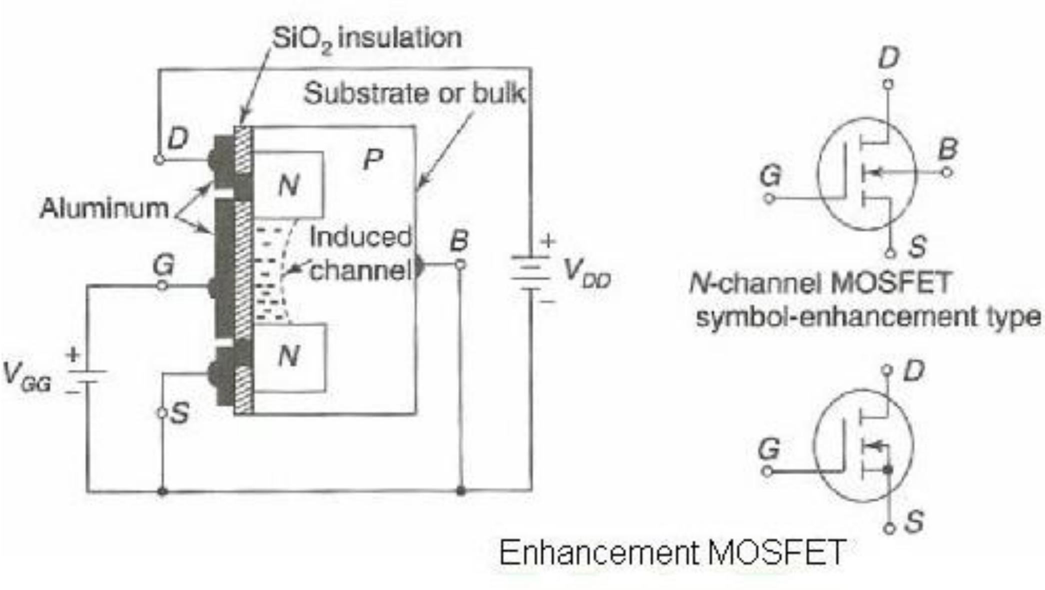


Figure: 1 Enhancement type MOSFET[1]

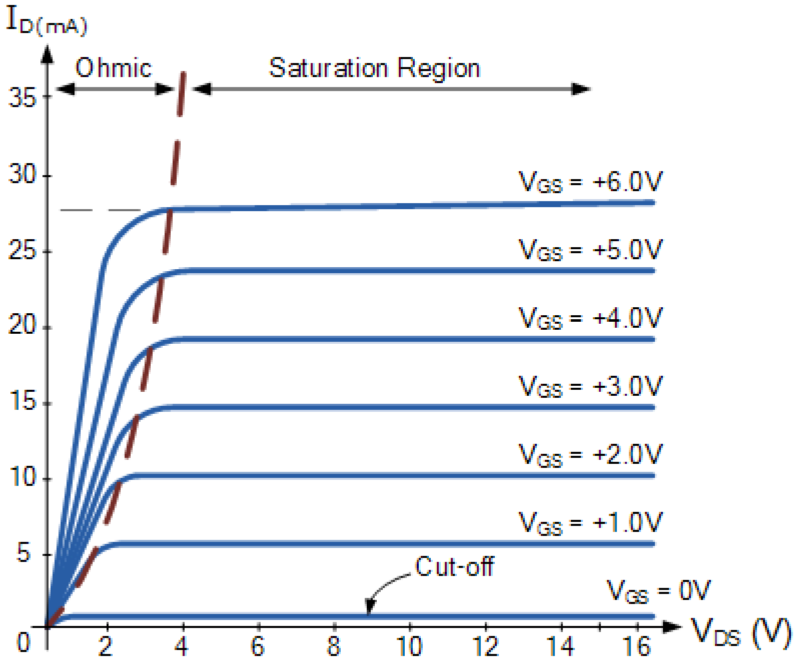


Figure 2: VDS vs ID curve[2]

The reverse is true for the p-channel enhancement MOS transistor. When VGS = 0 the device is “OFF” and the channel is open. The application of a negative (-ve) gate voltage to the p-type eMOSFET enhances the channels conductivity turning it “ON”. Then for an p-channel enhancement mode MOSFET: +VGS turns the transistor “OFF”, while -VGS turns the transistor “ON”.

Enhancement-mode MOSFETs make excellent electronics switches due to their low “ON” resistance and extremely high “OFF” resistance as well as their infinitely high input resistance due to their isolated gate. Enhancement-mode MOSFETs are used in integrated circuits to produce CMOS type Logic Gates and power switching circuits in the form of as PMOS (P-channel) and NMOS (N-channel) gates. CMOS actually stands for Complementary MOS meaning that the logic device has both PMOS and NMOS within its design.

p-channel Enhancement type MOSFET:

The p-channel is created when the transistor is properly biased. It is referred to as an enhancement mode p-channel MOSFET or PMOS.

Structure of PMOS:

The fabrication of a p-channel MOSFET starts with a substrate which is highly resistive n-type semiconductor. The base forms the body of the transistor. Diffused into the body are two low-resistivity p-type regions which are separated by the n-type substrate as it shown in the following figure.

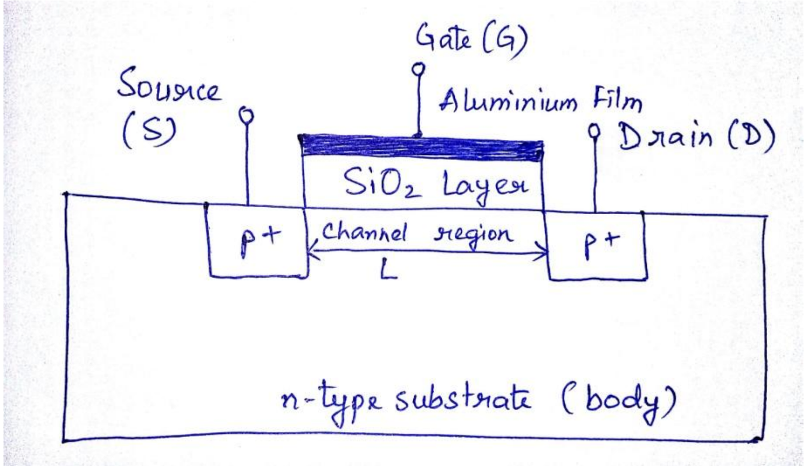


Figure3: Structure of a PMOS

Like the other FET, one of the p-regions is called the drain and the other is the source. For the gate to form a thin layer of SiO2 is grown over the surface of the n-type substrate. A thin film of aluminium film then acts as the gate terminal. As the device structure the p-channel is not yet created. The n-type substrate separates the two regions, there is very high resistance between the drain and the source. Unless it is properly biased the device is in it cut-off mode. Tat is why it is also referred to as a normally OFF device. Whereas the JFET and depletion type MOSFET are both normally ON device as the channel is present between the drain and the source terminals.

Induced p-channel in a PMOS:

We assume that the drain and the source terminals are held at a common potential. The gate terminal is insulated from the channel by the silicon dioxide layer, let us apply a negative voltage at the gate with respect to the source. The application of the negative voltage at the gate with respect to the source, VGS pulls the holes into the region between the drain and the source and pushes away the electrons. This is due to the capacitor formed by the insulating SiO2 layer between the gate and the p-type substrate. As the VGS is made of more negative, more and more holes are attracted towards the gate. These positive charges redistribute themselves in the region between the drain and the source in the form of thin layer. This layer of excess positive charges is called a p-type inversion layer. As VGS reaches a threshold voltage VT the entire region between the drain and the source gets filled with the positive charges. Thus a channel has been induced in the n region just below the gate. This channel connects the drain to the source. The conductivity region between the drain and the source increases and allows the current to flow from the source towards the drain when the drain is held at a lower potential than the source.

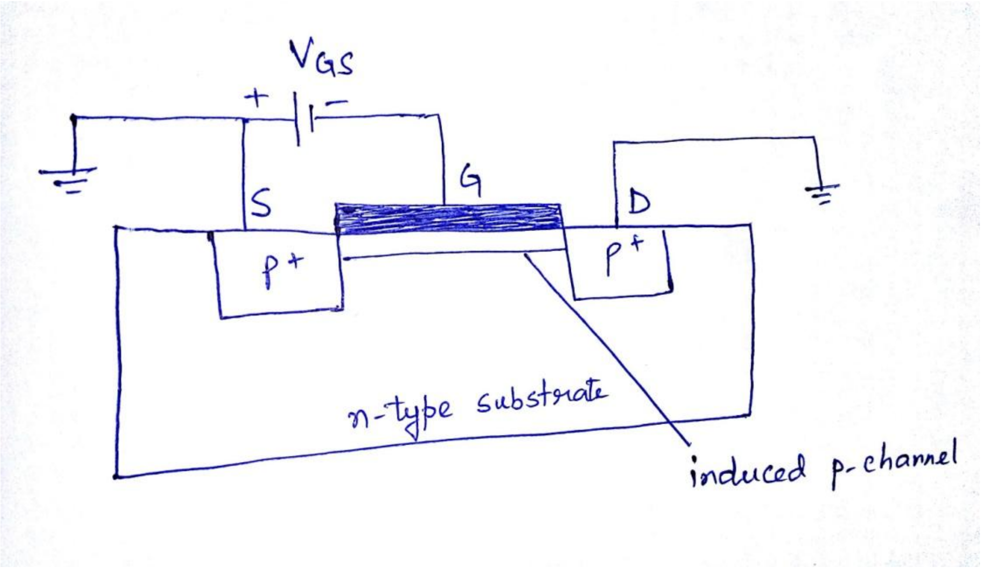


Figure 4: Induced channel in a PMOS

The drain current continues to increase with the decrease in the gate – to source voltage. Since the negative voltage increases the channel's conductivity. The PMOS is said to operate in its enhancement mode when the gate is negative with respect to the source. Therefore the PMOS operates in the enhancement mode when the source is positive with respect to the gate.

Channel Length Modulation:

If we apply a negative voltage at the drain with respect to the source VDS as shown in the following figure while the gate to source voltage VGS is less than VT is a negative voltage. As soon as VDS becomes less than zero there is a current, drain current from the source towards the drain through the highly conductive channel. With the decrease in VDS the gate to drain voltage increases which in turn results in the decrease in channel width near the drain terminal. The channel width near the source terminal remains the same as long as VGS is held constant. The decrease in channel width near the drain causes a decrease i its conductivity and an increase in its width near the drain causes a decrease in its conductivity and an increase in its resistance. As long as the channel exists between the drain and the source the drain current increases with the decrease in the drain –to source voltage. This is the linear/ triode region of operation of the PMOS.

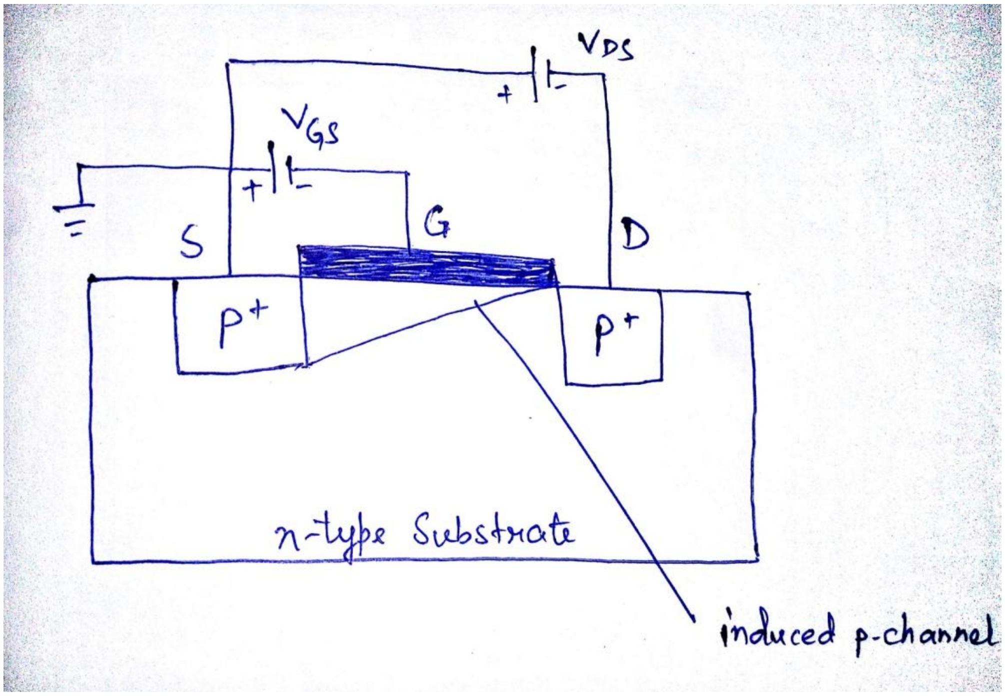


Figure 5: Channel length modulation of a PMOS

If we continue to decrease VDS the channel's width near the drain continues to diminish until it pinches off. This process is known as channel length modulation. At pinch –off, the drain to source is referred to as the channel modulation

Threshold Voltage:

The threshold voltage equals the sum of the flatband voltage, twice the bulk potential and the voltage across the oxide due to the depletion layer charge

$$V_T = V_{FB} + V_C + 2\phi_F + \frac{q(N_a - N_d)}{C_{OX}}\sqrt{\frac{2\varepsilon_S 2\phi_F}{q(N_a - N_d)}}$$

Where the Flatband voltage is given by:

$$V_{FB} = \Phi_{MS} - \frac{Q_f}{C_{OX}}$$

With

$$\Phi_{MS} = \Phi_M - \Phi_S = \Phi_M - (\chi + \frac{E_c - E_i}{2q} + \phi_F)$$

And

$$\phi_F = V_t \ln\frac{p}{n_i} = -V_t \ln\frac{n}{n_i}$$

The threshold voltage dependence on the doping density is illustrated with the figure below for both n-type and p-type MOS structures with an aluminum gate metal.

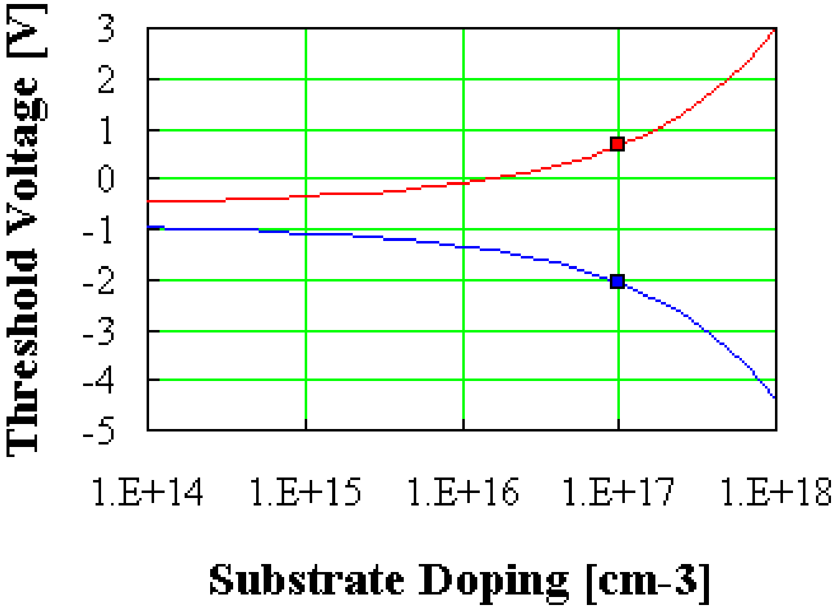


Figure 6: Variation of VT versus substrate doping

The thresholds of both types of devices is slightly negative at low doping densities and differs by 4 times the absolute value of the bulk potential. The threshold of nMOS capacitors increases with doping while the threshold of pMOS structures decreases with doping in the same way. A variation of the flatband voltage due to oxide charge will cause both curves to move down if the charge is positive and up if the charge is negative.

MOS CAPACITOR:

In a MOS capacitor, we replace the lower plate by a semiconductor. Unlike a metal, a semiconductor can have charges distributed in its bulk. For the sake of an example, let us consider a P type semiconductor (Si) doped to 1016 atoms /cm3 . As we know, holes outnumber electrons in this semiconductor by an extremely large factor. If we place a negative charge on the upper plate, holes will be attracted by this charge, and will accumulate near the silicon-insulator interface. This situation is analogous to the parallel plate capacitor and thus, the capacitance will be the same as that for a parallel plate capacitor. If, however, we place a positive charge on the upper plate, negative charges will be attracted by it and positive charges will be repelled. In a P type semiconductor, there are very few electrons. The negative charge is provided by the ionized acceptors after the holes have been pushed away from them. But the acceptors are fixed in their locations and cannot be driven to the edge of the insulator. Therefore, the distance between the induced and inducing charges increases - so the capacitance is lower as compared to the parallel plate capacitor. As more and more positive charge is placed on the upper plate, holes from a thicker slice of the semiconductor are driven away, and the incremental induced charge is farther from the inducing charge. Thus the capacitance continues to decrease. This does not, however, continue indefinitely. We know from the law of mass action that as hole-density reduces, the electron density increases. At some point, the hole-density is reduced and electron density increased to such an extent that electrons now become the "majority" carriers near the interface. This is called inversion. Beyond this point, more positive charge on the upper plate is answered by more electrons in the semiconductor. But the electrons are mobile, and will be attracted to the silicon insulator interface. Therefore, the capacitance quickly increases to the parallel plate value. Figure 3 shows the capacitance versus voltage (C-V) characteristics of a typical MOS capacitor.

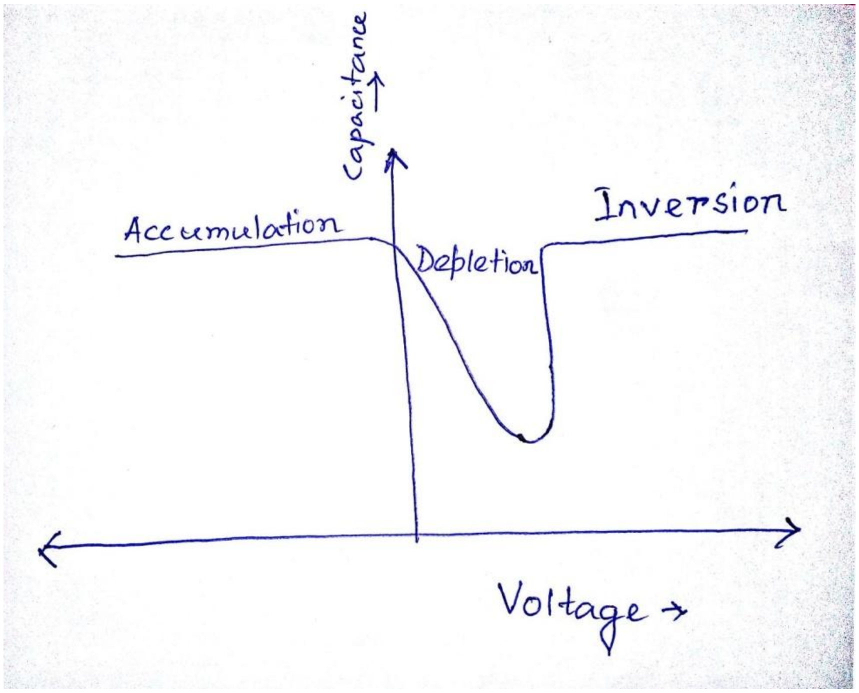


Figure 7: Curve showing the variation of capacitance with voltage

The induced interface charge in the MOS capacitor is closely linked to the shape of the electron energy bands of the semiconductor near the interface. At zero applied voltage, the bending of the energy bands is ideally determined by the difference in the work functions of the metal and the semiconductor. This band bending changes with the applied bias and the bands become flat when we apply the so-called flat-band voltage given by

$$V_{FB}= (\varphi_m – \varphi_s)/q= (\varphi_m – \varphi_s- E_c+ E_F)/q$$

Where φm and φs are work functions of metal and semiconductor respectively.

Xs is the electron affinity for the semiconductor, Ec is the energy of the conduction band edge, and EF is the Fermi level at zero applied voltage. The various energies involved are indicated in Figure 8, where we show typical band diagrams of a MOS capacitor at zero bias, and with the voltage V = VFB applied to the metal contact relative to the semiconductor–oxide interface.

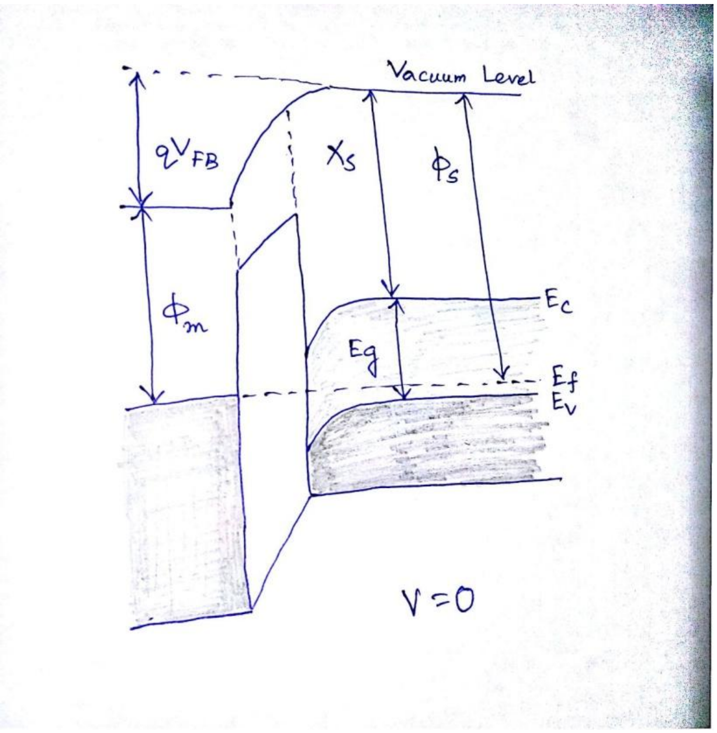


Figure 8: Band Diagrams of MOS Capacitor at zero bias

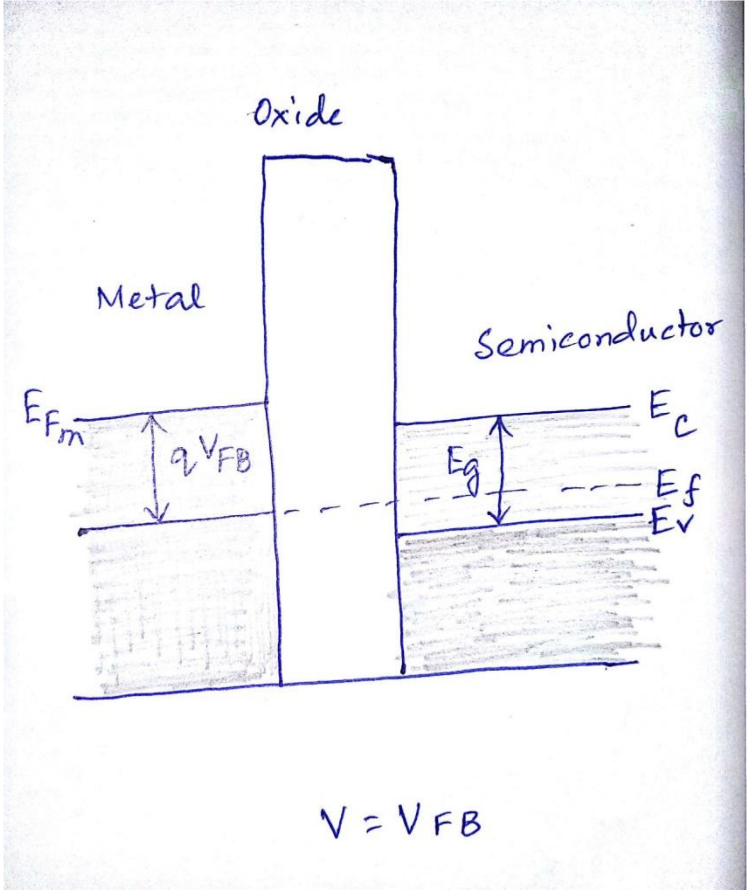


Figure 9: Band Diagrams of MOS Capacitor with an applied voltage equal to flat-band voltage

SMALL SIGNAL EQUIVALENT MODEL OF MOSFET FOR PI EQUIVALENT CIRCUIT:

The input resistance of this controlled source is high and we have assumed as it to be infinite upto now. In the analysis the FET can be replaced by the equivalent cicuit. The rest of the circuit remains unchanged except that ideal constant dc voltage sources are replaced by short circuits. This is aresult of the fact that the voltage across an ideal constant dc voltage source does not change and thus there will be a zero voltage signal across a constant dc voltage source.

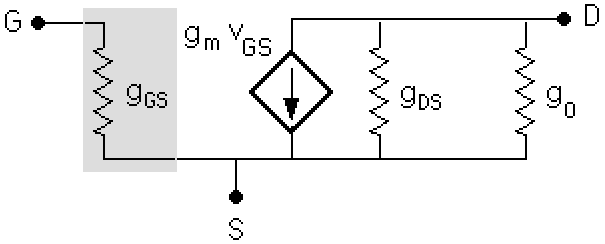


Figure: small signal model for MOSFET [3]

Reference:

1. http://teacher.buet.ac.bd/mjrahman/Class%20Note_Jellur.pdf
2.https://www.google.co.in/search?q=energy+band+diagram&source=lnms&tbm=isch&sa=X&ved=0ahUKEwiwrJ3E-pTUAhVEQo8KHQfND9MQ_AUIBigB&biw=1366&bih=638#imgrc=he-Z949fGKuK7M:
3. http://global-sei.com/sn/2012/420/5a.html
4. Streetman & Banerjee - Solid State Electronic Devices,PHI
5.S.M. Sze, Physics of semiconductor devices, John Wiley
6..http://www.ece.utep.edu/courses/ee3329/ee3329/Studyguide/ToC/Fundamentals/Carriers/explain.html
7.http://nptel.ac.in/courses/115102025/8
8.Callister, Materials Science and Engineering: An Introduction, Chapter 19.6-19.1
9. Milman, Halkias–Integrated Electronics – TMH
10.Sedra & Smith-Microelectronic Circuits- Oxford
11.Neamen- Semiconductor Physics and Devices TMH
12.S.M. Kang and Y. Leblebici. -CMOS Digital Integrated Circuits,Tata McGraw-Hill
13. https://www.physicsforums.com/threads/simple-derivation-of-diode-equation.307717/
14.http://www.rfwireless-world.com/Terminology/Depletion-MOSFET-vs-Enhancement-MOSFET.html
15.http://www.electronics-tutorials.ws/amplifier/mosfet-amplifier.html
16.http://people.seas.harvard.edu/~jones/es154/lectures/lecture_4/mosfet/mos_models/mos_models.html