Jaipur Engineering College & Research Centre, Jaipur

Department of Electronics & Communication Engineering

Session (2020-21)

Notes File

Electronics Devices (3EC4-07)

Prepared By:
Deepak sankhala,Mr.Mangilal Mr.Devdra Sharma
(AP, Dept. of ECE)

RAJASTHAN TECHNICAL UNIVERSITY, KOTA

SYLLABUS

II Year - III Semester: B.Tech. (Electronics & Communication Engineering)

3EC4-07: Electronic Devices

4 Credits | **Max. Marks: 200 (IA:40, ETE:160)**

3L:1T:0P | **End Term Exam: 3 Hours**

SN	Contents	Hours
1	Introduction to Semiconductor Physics: Introduction, Energy band gap structures of semiconductors, Classifications of semiconductors, Degenerate and non-degenerate semiconductors, Direct and indirect band gap semiconductors, Electronic properties of Silicon, Germanium, Compound Semiconductor, Gallium Arsenide, Gallium phosphide & Silicon carbide, Variation of semiconductor conductivity, resistance and bandgap with temperature and doping. Thermistors, Sensitors.	6
2	Review of Quantum Mechanics, Electrons in periodic Lattices, E-k diagrams. Energy bands in intrinsic and extrinsic silicon; Carrier transport: diffusion current, drift current, mobility and resistivity; sheet resistance, design of resistors.	6
3	Generation and recombination of carriers; Poisson and continuity equation P-N junction characteristics, I-V characteristics, and small signal switching models; Avalanche breakdown, Zener diode, Schottky diode.	8
4	Bipolar Junction Transistor, I-V characteristics, Ebers-Moll Model, MOS capacitor, C-V characteristics, MOSFET, I-V characteristics, and small signal models of MOS transistor, LED, photodiode and solar cell.	11
5	Integrated circuit fabrication process: oxidation, diffusion, ion implantation, Photolithography, etching, chemical vapor deposition, sputtering, twin-tub CMOS process.	9
	Total	**40**

UNIT 1 Introduction to Semiconductor Physics: Introduction, Energy band gap structures of semiconductors, Classifications of semiconductors, Degenerate and non-degenerate semiconductors, Direct and indirect band gap semiconductors, Electronic properties of Silicon, Germanium, Compound Semiconductor, Gallium Arsenide, Gallium phosphide & Silicon carbide,

Variation of semiconductor conductivity, resistance and bandgap with temperature and doping. Thermistors, Sensitors.

UNIT 2 Introduction to Semiconductor Physics: Introduction, Energy band gap structures of semiconductors, Classifications of semiconductors, Degenerate and non-degenerate semiconductors, Direct and indirect band gap semiconductors, Electronic properties of Silicon, Germanium, Compound Semiconductor,
Gallium Arsenide, Gallium phosphide & Silicon carbide, Variation of semiconductor conductivity, resistance and bandgap with temperature and doping. Thermistors, Sensitors.

UNIT 3 Generation and recombination of carriers; Poisson and continuity equation P-N junction characteristics, I-V characteristics, and small signal switching models; Avalanche breakdown,
Zener diode, Schottky diode

UNIT 4 Bipolar Junction Transistor, I-V characteristics, Ebers-Moll Model, MOS capacitor, C-V characteristics, MOSFET, I-V characteristics, and small signal models of MOS transistor, LED, photodiode and solar cell.

Unit 5 Integrated circuit fabrication process: oxidation, diffusion, ion implantation, Photolithography, etching, chemical vapor deposition, sputtering, twin-tub CMOS process.

UNIT- 1

INTRODUCTION TO SEMICONDUCTOR PHYSICS

INTRODUCTION: It has been proved by scientist that the operation of electronic devices depends upon the motion of charged particles within them .Therefore one should have an understading of the elements which control the motion of these particles. Certain materials are neither good conductor nor insulator ; their conductivity lies between conductor and insulator. We can classify materials by the energy gap between their valence band and the conduction band. The valence band is the band consisting of the valence electron, and the conduction band remains empty. Conduction takes place when an electron jumps from valence band to conduction band, and the gap between these two bands is forbidden energy gap. A wider the gap between the valence and conduction bands, higher the energy it requires for shifting an electron from valence band to the conduction band. In the case of **conductors**, this energy gap is absent or in other words conduction band, and valence band overlaps each other. Thus, electron requires minimum energy to jump from valence band. The typical examples of conductors are Silver, Copper, and Aluminium. In **insulators**, this gap is vast.

Therefore, it requires a significant amount of energy to shift an electron from valence to conduction band. Thus, insulators are poor conductors of electricity. Mica and Ceramic are the well-known examples of insulation material. **Semiconductors**, on the other hand, have an energy gap which is in between that of conductors and insulators.

This gap is typically more or less 1 eV, and thus, one electron requires energy more than conductors but less than insulating materials for shifting valence band to conduction band. At low temperature there are very less number of electrons in conduction band in a semiconductor crystal but when the temperature is increased more and more electrons get sufficient energy to migrate from valence band to conduction band. Because of that, they don't conduct electricity at low temperature but as the temperature increases the conductivity increases. The most typical examples of the semiconductors are silicon and germanium.

Definition of Semiconductor : A semiconductor material is one whose electrical properties lie in between those of insulators and good conductors. Examples are: germanium and silicon. In terms

of energy bands, semiconductors can be defined as those materials which have almost an empty conduction band and almost filled valence band with a very narrow energy gap (of the order of 1 eV) separating the two.

Let us list out some of the key points of semiconductor applications:

- The Wide choice to alter the physical properties (example: high to low electrical resistivity,optical materials covering from UV to far infrared applications).
- Ultra-fast response times .
- Multi-functional (source, operation and detection) devices
- Tailoring opportunity, in such a way that the device allows to implement all sorts of information processing.
- Boolean logics, signal amplifiers, store and retrieve information, processing, light source… many more.

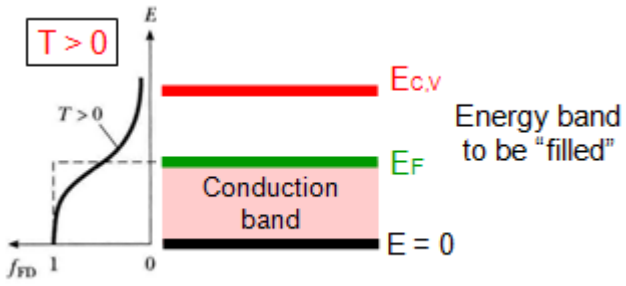
There are three types of material which are generally studied.

these are matel,insulator and semiconductor.

Classification of materials :

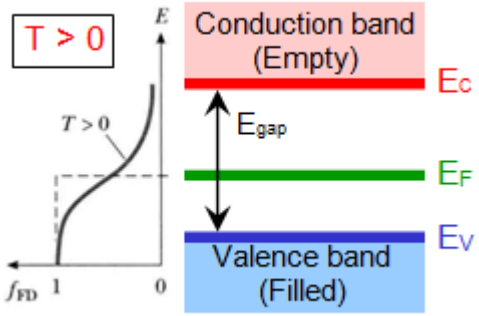
Metals:

Highest occupied energy bands are partially filled with electrons, while above the Fermi energy level(E_F) all bands are empty. With a very small amount of energy lead the electrons go to the conduction band, leading to high conductivity. At T>0 electrons thermally excited and cross the barrier of E_F .



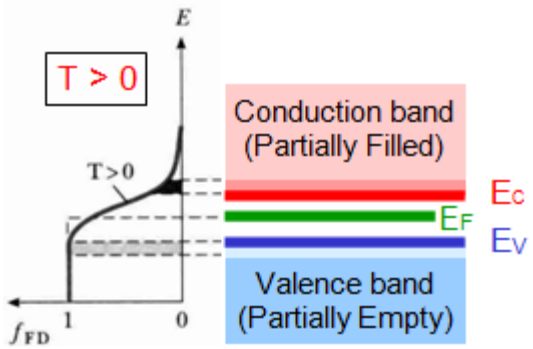
Insulators:

At T=0 Conduction band is completely empty and valence band is filled, leading to zero conductivity. Very big energy gap (E_g>4,5 eV)between conduction (E_C) and Valance (E_V) bands, where E_F is in the middle. No thermal excitation and hence even higher temperature, conduction are zero.



Semiconductors:

At T=0, conduction band is empty and valence is completely filled, hence zero conductivity. E_g is < 4eV(Si: 1.17, Ge=0.74, GaAs= 1.52eV at T=0). At higher temperature electron thermal excitation does happen and hence the conductivity lies between metal and insulator.



Metal	Semiconductor	Insulator
Silver, platinum copper gold.. etc	Silicon, Germanium, etc.,	Teflon, Quartz, SiO_2 etc.,
Conductivity : 10^6 to 1 $(\Omega\text{-m})^{-1}$	10^{-8} to 10^3	10^{-7}- 10^{-25}
Highest occupied energy bands are partially filled with electrons:	Completely filled at low temperatures. Upper band conduction band is empty	

Table 1.1

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

In general, any material property varies typically according to the electronic configuration. If we look into the periodic table, the group elements in a column have similar properties, gradually changing from top to bottom. Mostly, the electrons are in the outmost shell govern most of the physical properties. Our Semiconductors occupy some of the III to IV columns in the periodic table and they could be a mix of various combinations called “alloys”.

One of the uniqueness of semiconductor is that the properties can be widely tailored, according to our requirement, by simply modifying the alloy combinations. Apart from this, we can also add a small fraction of other materials, so called doping, and change the properties. The best example is doping in Si. We can change the property of Silicon to electron rich (n-type) or deficient (p-type) by adding(doping) Boron or Phosphorous atoms respectively.

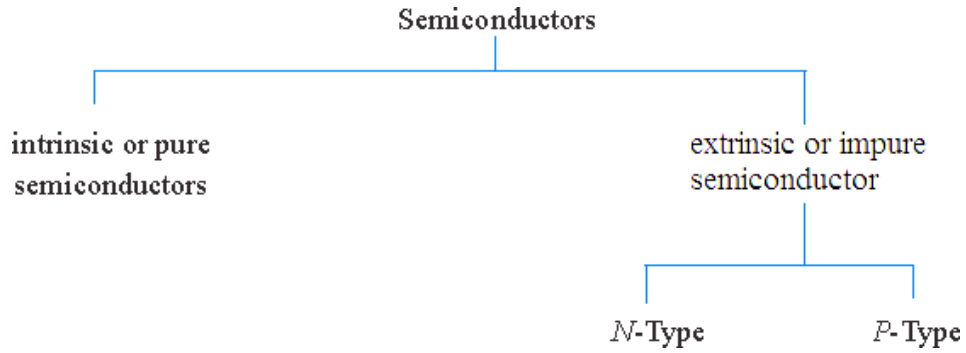
Some of the important Semiconductors and their combinations (alloys) are

Table 1.2

Elemental	**IV alloys**	**III-V binary alloys**	**II-VI binary**	**Ternary alloys**
Si, Ge	**SiC,** **SiGe**	**Al Y (Y=P,As,Sb),** **GaY(Y=N,P,As,Sb)** **IY(Y=P,As,Sb)**	**ZnX,** **CdX** **(X=S, Se and Te)**	**AlGaAs,**

Another important aspect in learning these material characteristics is how atoms are arranged. They can be either crystalline (well-ordered), amorphous (disordered) or polycrystalline (short range order).

Semiconductor may be classified as under:



N-type (e.g. doped with Antimony)

These are materials which have **Pentavalent** impurity atoms (Donors) added and conduct by "electron" movement and are therefore called, **N-type Semiconductors**.

In N-type semiconductors there are:

- 1. The Donors are positively charged.
- 2. There are a large number of free electrons.
- 3. A small number of holes in relation to the number of free electrons.
- 4. Doping gives:
    - positively charged donors.
    - negatively charged free electrons.
- 5. Supply of energy gives:
    - negatively charged free electrons.
    - positively charged holes.

P-type (e.g. doped with Boron)

These are materials which have **Trivalent** impurity atoms (Acceptors) added and conduct by “hole” movement and are therefore called, **P-type Semiconductors**.

In these types of materials are:

- 1. The Acceptors are negatively charged.
- 2. There are a large number of holes.
- 3. A small number of free electrons in relation to the number of holes.
- 4. Doping gives:
    - negatively charged acceptors.
    - positively charged holes.
- 5. Supply of energy gives:
    - positively charged holes.
    - negatively charged free electrons.

and both P and N-types as a whole, are electrically neutral on their own.

Antimony (Sb) and Boron (B) are two of the most commonly used doping agents as they are more feely available compared to other types of materials. They are also classed as “metalloids”. However, the periodic table groups together a number of other different chemical elements all with either three, or five electrons in their outermost orbital shell making them suitable as a doping material.

These other chemical elements can also be used as doping agents to a base material of either Silicon (Si) or Germanium (Ge) to produce different types of basic semiconductor materials for use in electronic semiconductor components, microprocessor and solar cell applications. These additional semiconductor materials are given below.

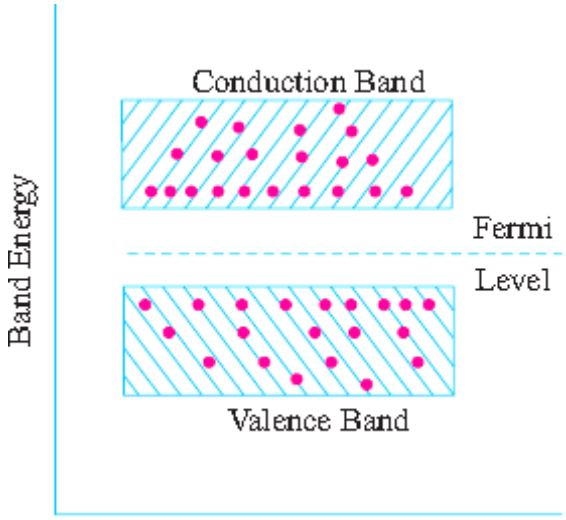
Insic Semiconductors :

An intrinsic semiconductor is one which is made of the semiconductor material in its extremely pure form.

Examples of such semiconductors are: pure germanium and silicon which have forbidden energy gaps of 0.72 eV and 1.1 eV respectively. The energy gap is so small that even at ordinary room temperature; there are many electrons which possess sufficient energy to jump across the small energy gap between the valence and the conduction bands.

Alternatively, an intrinsic semiconductor may be defined as one in which the number of conduction electrons is equal to the number of holes.

Schematic energy band diagram of an intrinsic semiconductor at room temperature is shown in Fig. below.



Extrinsic Semiconductors:

Those intrinsic semiconductors to which some suitable impurity or doping agent or doping has been added in extremely small amounts (about 1 part in 10^8) are called extrinsic or impurity semiconductors.

Depending on the type of doping material used, extrinsic semiconductors can be sub-divided into two classes:

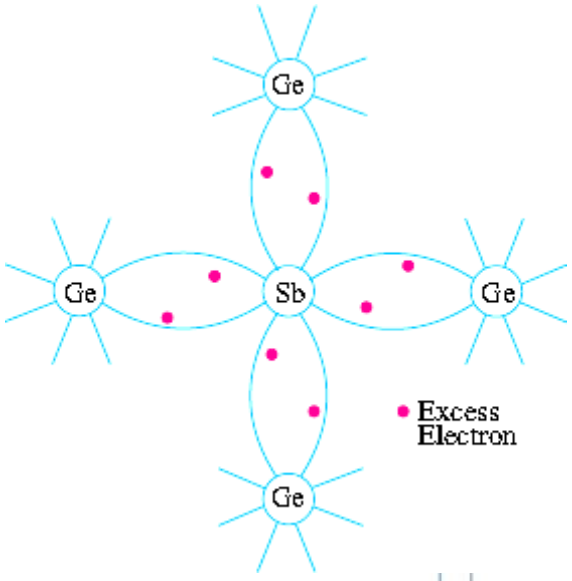
(i) N-type semiconductors and

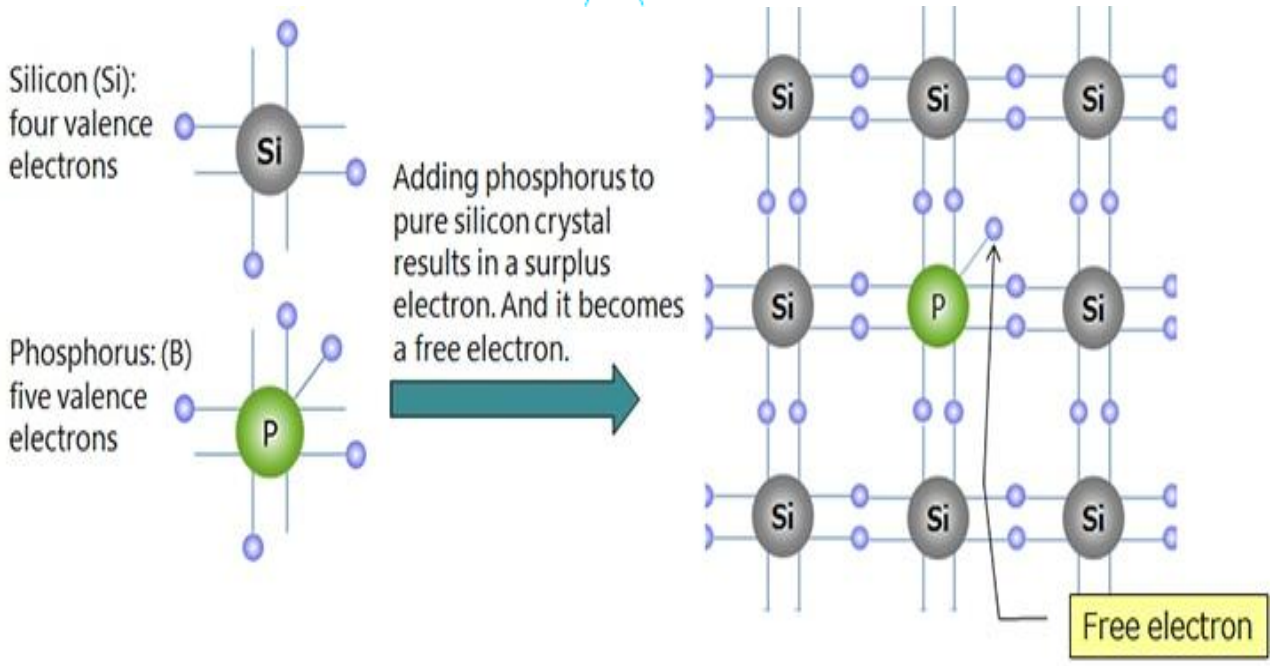
(ii) P-type semiconductors.

(i) N-type Extrinsic Semiconductor:

This type of semiconductor is obtained when a pentavalent material like antimonty (Sb) is added to pure germanium crystal. As shown in Fig. below, each antimony atom forms covalent bonds with the surrounding four germanium atoms with the help of four of its five electrons. The fifth electron is superfluous and is loosely bound to the antimony atom.

Hence, it can be easily excited from the valence band to the conduction band by the application of electric field or increase in thermal energy. It is seen from the above description that in N-type semiconductors, electrons are the majority carriers while holes constitute the minority carriers.



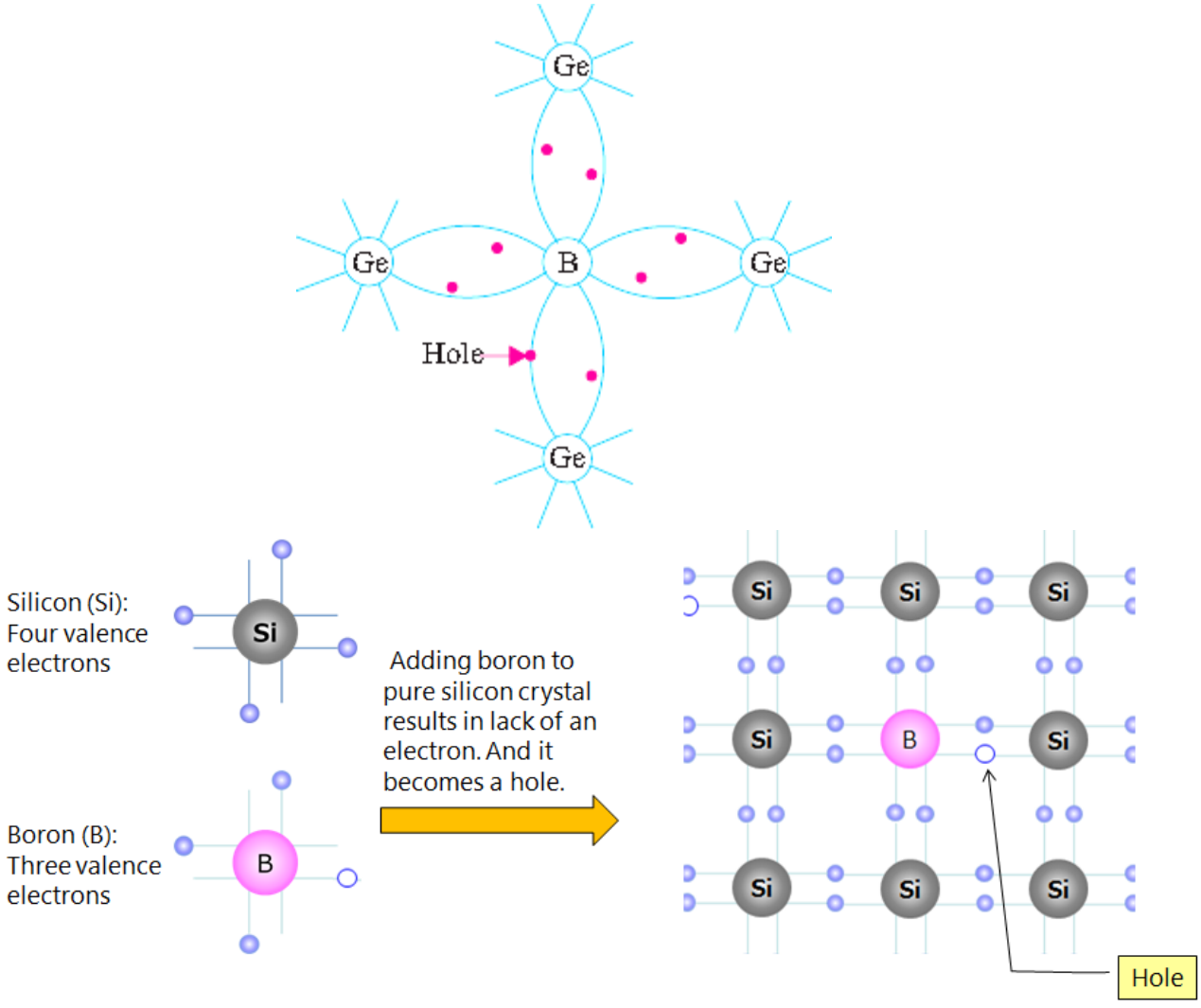


(ii) P-type Extrinsic Semiconductor:

This type of semiconductor is obtained when traces of a trivalent like boron (B) are added to a pure germanium crystal. In this case, the three valence electrons of boron atom form covalent bonds with four surrounding germanium atoms but one bond is left incomplete and gives rise to a hole as shown in Fig. below.

Thus, boron which is called an acceptor impurity causes as many positive holes in a germanium crystal as there are boron atoms thereby producing a P-type (P for positive) extrinsic semiconductor.

In this type of semiconductor, conduction is by the movement of holes in the valence band.



Majority and Minority Carriers:

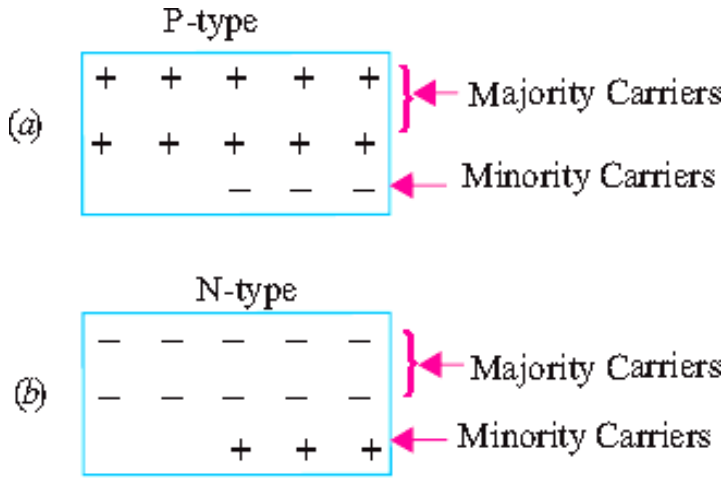
In a piece of pure germanium or silicon, no free charge carriers are available at 0°K. However, as its temperature is raised to room temperature, some of the covalent bonds are broken by heat energy and as a result, electron-hole pairs

Produced. These are called thermally-generated charge carriers. They are also known as intrinsically-available charge carriers. Ordinarily, their number is quite small. An intrinsic of pure germanium can be converted into a P-type semiconductor by the addition of an acceptor impurity which adds a large number of holes to it. Hence, a P-type material contains following charge carriers:

(a) Large number of positive holes—most of them being the added impurity holes with only a very small number of thermally generated ones.

(b) A very small number of thermally-generated electrons (the companions of the thermally generated holes mentioned above).

Obviously, in a P-type material, the number of holes (both added and thermally- generated) is much more than that of electrons. Hence, in such a material, holes constitute majority carriers and electrons form minority carriers as shown in Fig. below (a). Similarly, in an N-type material, the number of electrons (both added and thermally-generated) is much larger than the number of thermally-generated holes. Hence, in such a material, electrons are majority carriers whereas holes are minority carriers as shown in Fig. below (b).



Types of Non-degenerate and degenerate semiconductor :

Non-degenerate semiconductors contain moderate level of doping, where dopant atoms are well separated from each other with negligible interactions. Consequently, the dopant atoms exhibit discrete energy levels and are usually formed either below the conduction band edge or on top of the valence band edge.

Degenerate:

1. Degenerate semiconductors contain high level of doping, with significant interaction between dopant atoms. The interaction results in the formation of donor/acceptor bands rather than discrete energy levels. These impurity bands can overlap with the corresponding band edges of conduction or valence bands.
2. At such high doping levels the Fermi level is expected to lie in the conduction or valence bands rather than in the forbidden region.
3. The degenerate semiconductors can act like a metal if the Fermi level lies in the conduction band even at absolute zero K (usually observed in metals). In other words, the electrical conductivity decreases with the increase in temperature. If the Fermi level lies outside at low temperatures the degenerate semiconductor behaves like a semiconductor.

Degenerative bandgaps are heavily dopped and bandgap simultaneously shrinks and the Fermi levels Ef either shifts above Ec or below Ev.

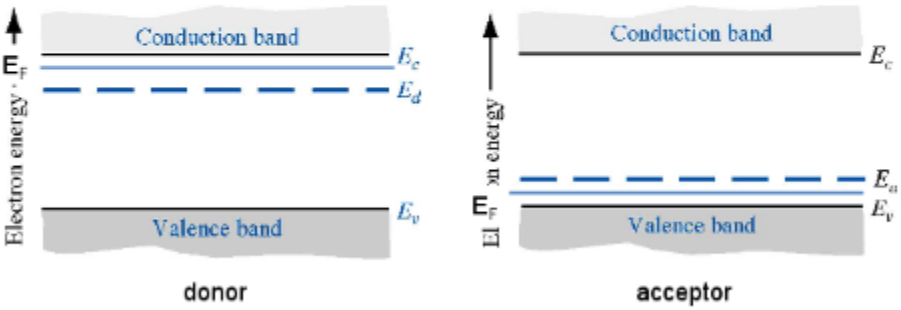
The non-degenerative is just opposite as it is lightly doped and the Fermi level Ef is 3kT above Ev valence band or 3kT below Ec conduction band. Where k is the boltzman constant and T is the temperature in kelvin.

Degenerate semiconductors has a high level containing doping with significant interaction between dopants. while non degenerate semiconductors contain moderate level of doping, where the dopant atoms are well separated from each other in the semiconductor host lattice with negligible interaction

Degenerate and Non degenerate semiconductors

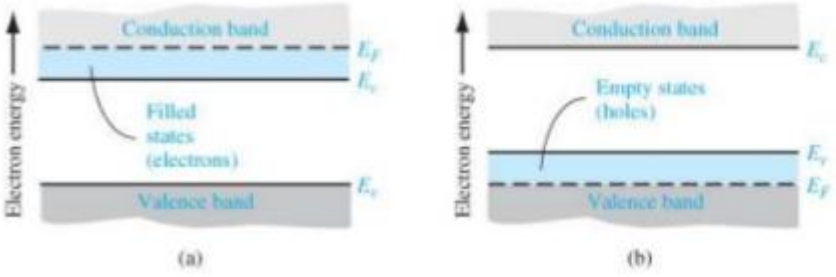
Small amount of dopant atoms (impurity atoms)

- ✓No interaction between dopant atoms
- ✓Discrete, noninteracting energy state.
- ✓E_F at the bandgap



Nondegenerate semiconductor

- As the donor concentration further increases, the band of donor states widens and may overlap the bottom of the conduction band.
- This overlap occurs when the donor concentration becomes comparable with the effective density of states.
- **When the concentration of electrons in the conduction band exceeds the density of states N_c**, the Fermi energy lies within the conduction band. This type of semiconductor is called a **degenerate** n-type semiconductor.
- In the degenerate n-type semiconductor, **the states between E_f and E_c are mostly filled with electrons**; thus, the electron concentration in the conduction band is very large.

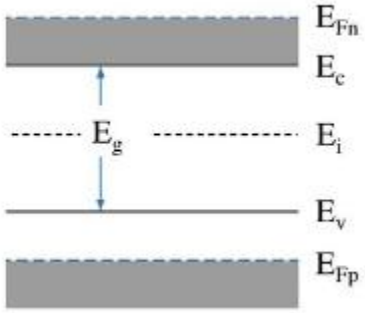


- if the concentration of electrons in the conduction band, n, exceeds the effective density of states, N_c

  - then the Fermi level is no longer within the band gap

  - it lies somewhere in the conduction band

  - when this occurs, the material is called degenerate n-type

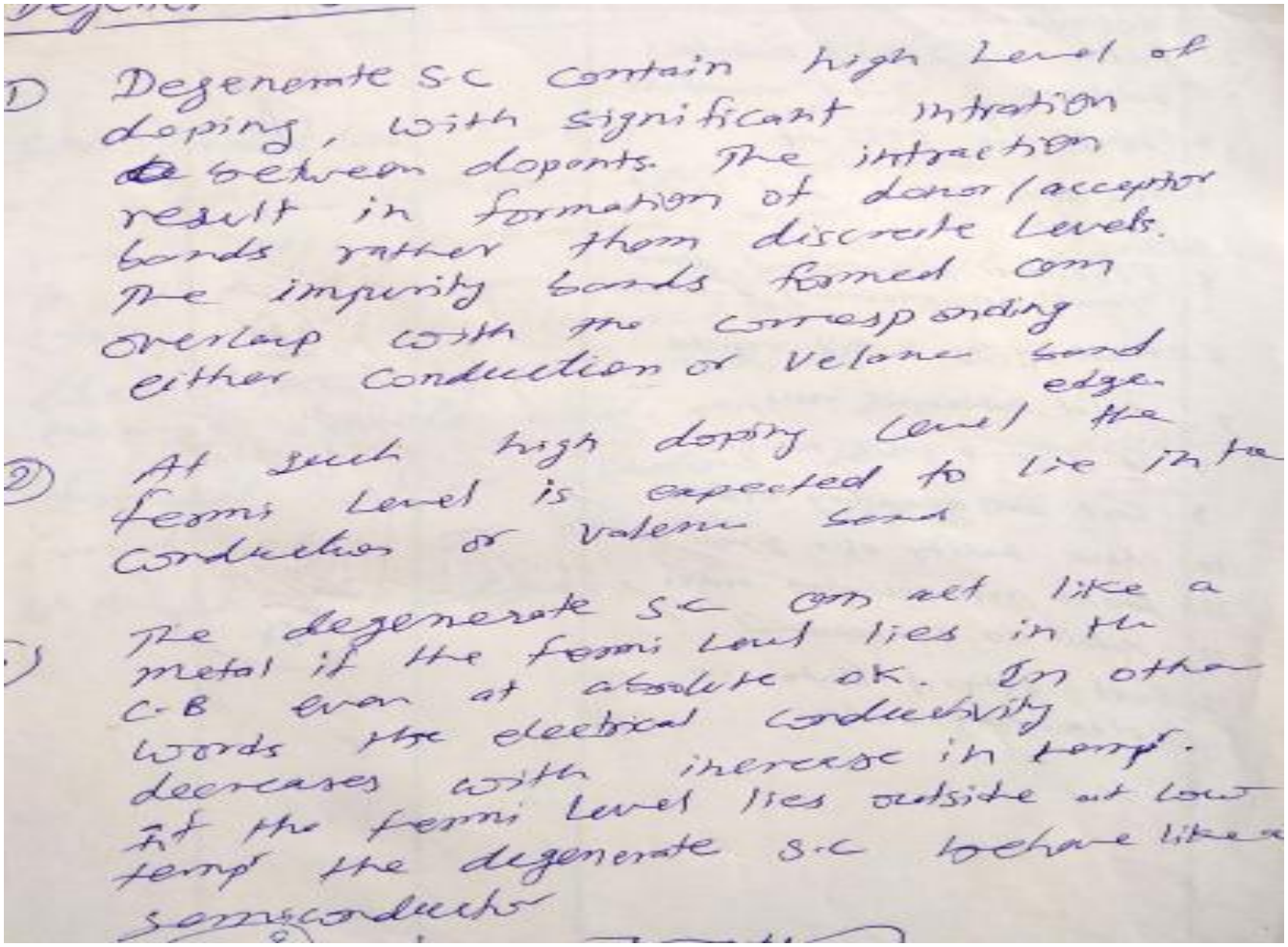


- an equivalent situation exists for the case where the acceptor impurity concentration is large, in this case, the material is called degenerate p-type, and the Fermi level lies within the valence band

- energy states below E_F are mostly filled, and states above E_F are mostly empty

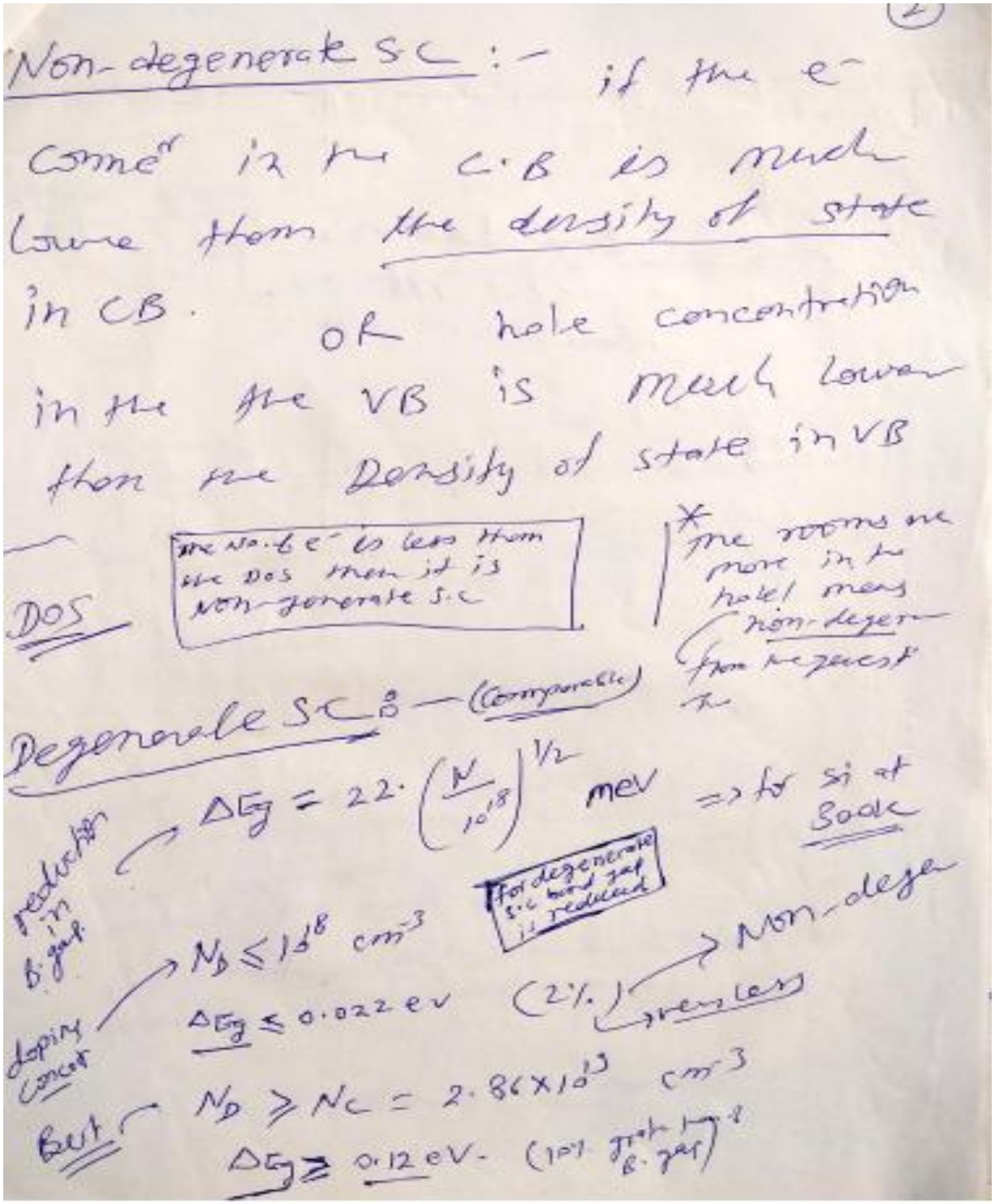
  - in a degenerate n-type material, the region between E_c and E_F is mostly filled with electrons

  - in a degenerate p-type material, the region between E_v and E_F is mostly filled with holes


1) Degenerate S.C contain high Level of doping, with significant intration between dopants. The intraction result in formation of donor/acceptor bands rather than discrete Levels. The impurity bands formed can overlap with the corresponding either conduction or velance band edge.

2) At such high doping Level the fermi Level is expected to lie in the conduction or valence band

3) The degenerate S.C can act like a metal if the fermi Level lies in the C.B even at absolute 0K. In other words the electrical conductivity decreases with increase in temp. If the fermi Level lies outside at Low temp the degenerate S.C behave like a semiconductor


Non-degenerate S.C :- if the e⁻ conc^n in the C.B is much lower than the density of state in CB. OR hole concentration in the the VB is much lower than the Density of state in VB
DOS
the No. of e⁻ is less than the DOS than it is Non-degenerate S.C
* the rooms are more in the hotel means non-degen from request to
Degenerate S.C :- (Compared)
predictor in B.gap
$\Delta E_g = 22 \cdot \left(\frac{N}{10^{18}}\right)^{1/2}$ meV
=> for Si at 300K
For degenerate S.C band gap is reduced
doping concn
$N_D \leq 10^{18}$ cm^{-3}
$\Delta E_g \leq 0.022$ eV (2%)
Non-dege
negligible
But
$N_D \geq N_c = 2.8 \times 10^{19}$ cm^{-3}
$\Delta E_g \geq 0.12$ eV. (10% of B.gap)

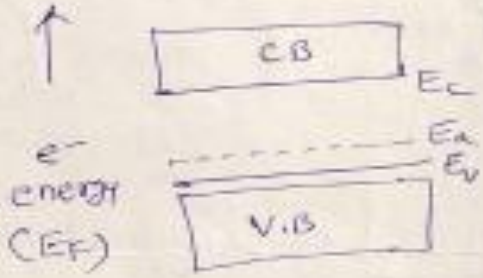
Degenerate and Non-degenerate Semiconductors:-

✓ Small amount of dopant (Impurity atoms)

1. No interaction between dopant atoms
2. Discrete, noninteracting energy state
3. E_F at the bandgap

Donor

(Non-degenerate S-c)



Large Amount of dopant Atoms:-

(N_c effective density of states)

1. Dopant Atoms interact with each other
2. Band of dopant states widens and overlap the allowed band (conduction @ valence band)
3. E_F lies within conduction @ Valence band

Non-Degenerate S.C :-

If the electron concentration in the Conduction band is much lower than the density of states in the conduction band

OR

Hole concentration in the Valence band is much lower than the density of states in the Valence band

→ What is the value of →

→ For Degenerate S.C the E gap reduce

$$\Delta E_g = 22 \cdot \left(\frac{N}{10^{18}}\right)^{1/2} \text{ meV}$$

N → No. of e⁻ or holes

reduction in band gap

for Si at 300k

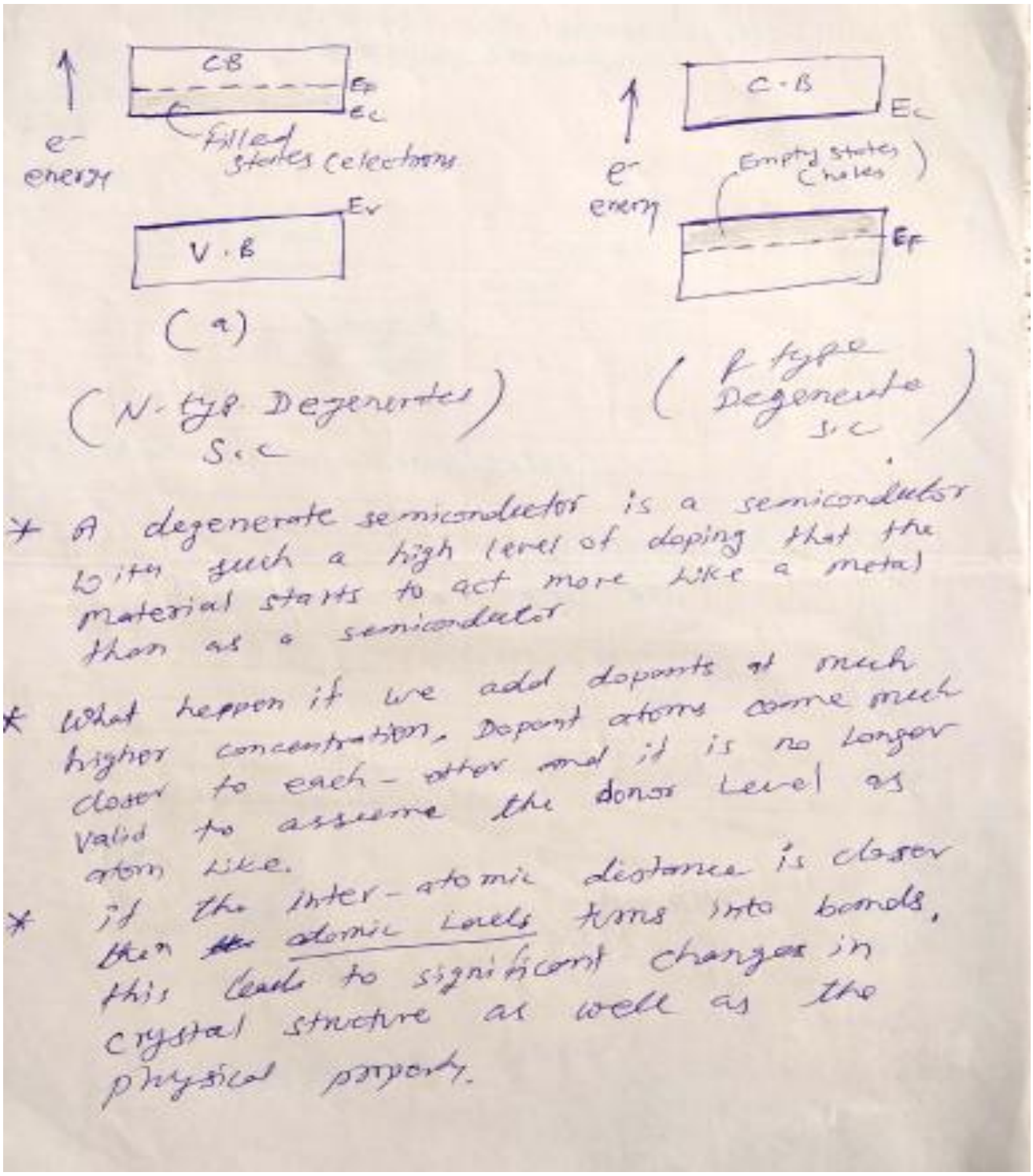
doping Level

for $N_D \leq 10^{18}$ cm^{-3} — Non degen

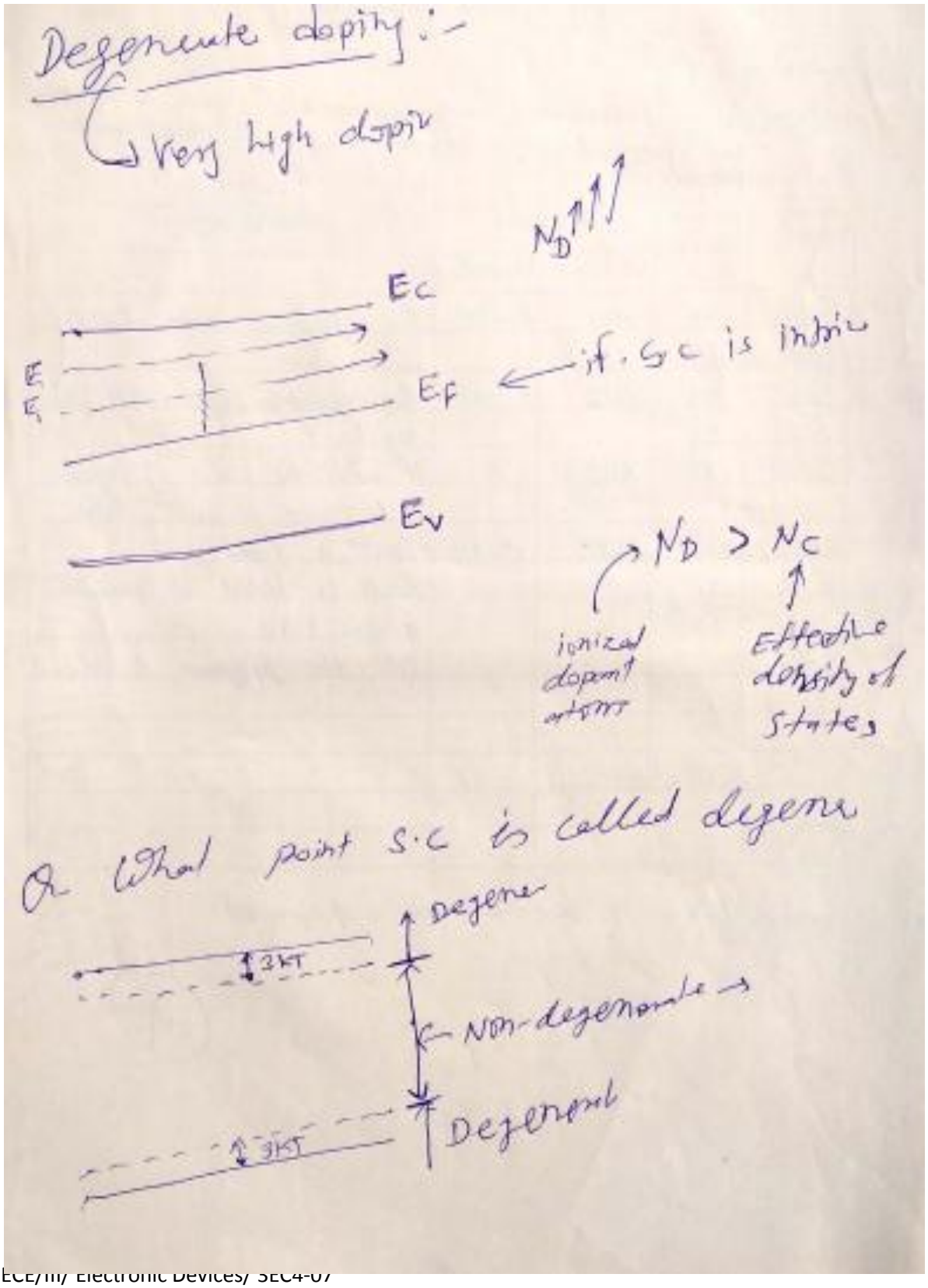
$\Delta E_g \leq 0.022$ eV ($\approx 2\%$)

for $N_D \geq N_C = 2.8 \times 10^{19}$ cm^{-3} — degen

$\Delta E_g \geq 0.12$ eV (10%)



* A degenerate semiconductor is a semiconductor with such a high level of doping that the material starts to act more like a metal than as a semiconductor
* What happen if we add dopants at much higher concentration, Dopant atoms come much closer to each-other and it is no longer valid to assume the donor Level as atom like.
* if the inter-atomic distance is closer then ~~the~~ atomic Levels turns into bands, this leads to significant changes in crystal structure as well as the physical property.


Degenerate doping :-
Very high doping
N_D
E_c
E_f
E_v
if S.C is intrinsic
$N_D > N_C$
ionized dopant atoms
Effective density of states
At what point S.C is called degen
Degener
3KT
Non-degenerate
Degenerat
3KT

Degenerate and Non-Degenerate S-C :-

Density of state is a conspet where you can get the no of available state of e or h in the material

Level → energy Level
or
state

No. of state are known as density of state

How many Levels are there

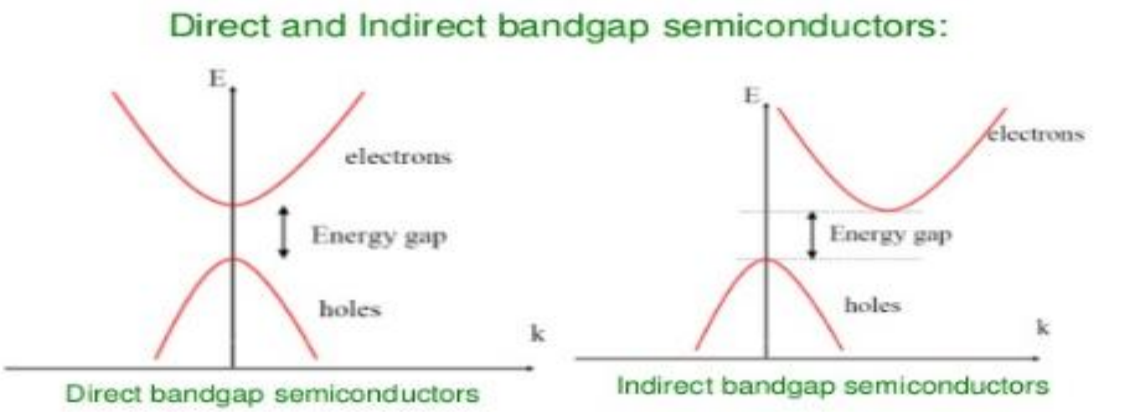
Level means where e or hole are accumulate

DIRECT AND INDIRECT BAND GAP SEMICONDUCTORS

In a direct band gap semiconductor, the top of the valence band and the bottom of the conduction band occur at the same value of momentum. In an indirect band gap semiconductor, the maximum energy of the valence band occurs at a different value of momentum to the minimum in the conduction band energy.

1. Direct and Indirect bandgap semiconductors: Those materials for which maximum of valence band and minimum of conduction band lie for same value of k, called direct bandgap materials (i.e. satisfies the condition of energy and momentum conservation). For example: GaAs, InP, CdS..etc Direct bandgap semiconductors Indirect bandgap semiconductors Those materials for which maximum of valence band and minimum of conduction band do not occur at same value of k, called indirect bandgap materials. For example: Si and Ge
2. Indirect bandgap materials are not suitable for optical devices (LEDs and Laser diodes) Direct bandgap semiconductors: Indirect bandgap semiconductors:
3. Direct bandgap semiconductors A common and simple method for determining whether a band gap is direct or indirect uses absorption spectroscopy. By plotting certain powers of the absorption coefficient against photon energy, one can normally tell both what value the band gap has, and whether or not it is direct. $\alpha 2$ n is the (real) index of refraction xvc is a "matrix element", with units of length and typical value the same order of magnitude as the lattice constant. if a plot of $h\nu$ versus $\alpha 2$ forms a straight line, it can normally be inferred that there is a direct band gap, measurable by extrapolating the straight line to the $\alpha = 0$ axis. $h\upsilon$
4. Indirect bandgap semiconductors Ep is the energy of the phonon that assists in the transition $\alpha 1/2$ $h\upsilon$ Ep is the energy of the phonon that assists in the transition if a plot of $h\nu$ versus $\alpha 1/2$ forms a straight line, it can normally be inferred that there is a indirect band gap, measurable by extrapolating the straight line to the $\alpha = 0$ axis (assuming Ep=0).
5. 1D confinement: quantum wells; structures consisting of a thin well materials sandwiched between two layers of a barrier materials. 2D confinement: quantum wires; structures consisting of a thin and narrow well materials surrounded by barrier materials. 3D confinement: quantum dots; nano-size particles in a barrier materials. Why Nanomaterials ? materials. The quantum confinement => allowed electron and

hole states are quantized in the well region => energy required to generate e-h pair or radiation emitted from the process of e-h pair recombination is modified => wavelength tuning of the radiation (used in LED or laser applications)



➢ Those materials for which maximum of valence band and minimum of conduction band lie for same value of k, called direct bandgap materials (i.e. satisfies the condition of energy and momentum conservation). For example: GaAs, InP, CdS..etc

➢ Those materials for which maximum of valence band and minimum of conduction band do not occur at same value of k, called indirect bandgap materials. For example: Si and Ge

In a direct band gap semiconductor, the top of the valence band and the bottom of the conduction band occur at the same value of momentum.

In an indirect band gap semiconductor, the maximum energy of the valence band occurs at a different value of momentum to the minimum in the conduction band energy.

For direct or indirect identification you must first calculate the band gap with the formula 1240/λ (λ is absorption edge that obtained from the DRS analysis) and then compare with each of the numbers achieved with the method that Dipayan Pal has mentioned.

In semiconductor physics, the band gap of a semiconductor is always one of two types, a direct band gap or an indirect band gap. The minimal-energy state in the conduction band and the maximal-energy state in the valence band are each characterized by a certain crystal momentum (k-vector) in the Brillouin zone. If the k-vectors are

the same, it is called a "direct gap".If they are different, it is called an "indirect gap". The band gap is called "direct" if the momentum of electrons and holes is the same in both the conduction band and the valence band; an electron can directly emit a photon. In an "indirect" gap, a photon cannot be emitted because the electron must pass through an intermediate state and transfer momentum to the crystal lattice.

ELECTRONIC PROPERTIES OF VARIOUS ELECTRONIC MATERIALS

1.31 Electronic properties of different Semiconductor Materials

The properties and applications of generally used semiconducting materials are as follows :

1. Germanium

(a) Melting point : 937°C

(b) Atomic number : 32

(c) Energy gap : 0.67eV

(d) It is a grey metallic looking material.

(e) It is brittle and glass like in its mechanical properties.

(f) It crystallizes in the diamond cubic lattice.

(g) It has an intrinsic resistivity of 46 Ω–cm but may be doped with Antimony or Arsenic to give N–type resistivity of 0.01 Ω–m or less and with Boron, Gallium or Aluminium to give P–type resistivity of 0.001 Ω–cm or less.

(h) It is usually prepared in high purity single crystal form for electronics used by pulling from the melt, either vertically or horizontally. To obtain intrinsic material, impurities of active elements must be in the range of 1 part/billion or less.

(i) The active impurities such as Gallium, Indium, Arsenic and Antimony tend to diffuse rather slowly in Germanium. The donors diffuse faster than the acceptors.

(j) It finds use in near infrared detection and in x–ray detectors.

2. Silicon

(a) Melting point : 1410°C

(b) Atomic Number : 14

(c) Energy gap : 1.1eV

(d) Silicon is handicapped by the lower carrier mobilities 1350 and 980 cm^2/V–sec for electrons and holes respectively compared with values of 3900 and 1800 cm^2/Vsec for Germanium. This drawback limits applications of silicon in high frequency transistors.

(e) The properties of silicon are notably sensitive to the presence of oxygen. Oxygen in silicon tends to introduce instabilities when the material is subjected to heat treatments at high temperature.

(f) Silicon is one of the most sensitive element to nuclear radiations.

(g) Silicon is doped with P, As and Sb donors and B, Al, Ga acceptors.

(h) The bulk properties of Silicon under radiation are particularly sensitive to the presence of oxygen.

(i) It is used for diodes, transistors, ICs, infrared detection, photosensors, solar cells, charge coupled devices and variety of other devices.

3. Gallium Arsenide

(a) Energy gap : 1.4eV at room temperature.

(b) Its characteristics are very closely related to those of Germanium. This is under stable, since Gallium is the third column neighbour and arsenic is the fifth column neighbour of Germanium in the periodic table.

(c) The melting point is 1300°C combined with the high vapour pressure of arsenic at 1200°C to 1300°C makes the production of Gallium Arsenide of electronic grade an extremely difficult one.

(d) Horizontal furnances using zone refining technique have proved the best but the resulting material still is not competitive with silicon and germanium in the purity or structural perfection.

(e) It is used for Schottky diodes, light emitting diodes, gun diodes and lasers.

4. Indium Antimonide

(a) Melting point : 525°C

(b) It has small band gap and very high mobility, i.e., around 70000 cm^2/V–sec.

(c) It is much easier to prepare in single crystal form than Gallium Arsenide.

(d) It is useful for infrared detectors, infrared filter material and hall effect devices. Transistors, Tunnel diodes and Laser diodes are also manufactured from Indium Antimonide (InSb).

5. Cadmium Sulphide

(a) Energy gap : 2.4eV

(b) The colour of pure CdS is pale yellow.

(c) It melts only under at high pressure.

(d) It can be prepared in the resistivity range from 10 to about 10^{12} Ω–cm depending on the presence of defects and the impurities.

(e) It can be prepared as an n–type semiconductor and is extremely difficult to produce p–type CdS.

(f) It has been used commerically as a photo conductor for many years as well as constituent of cathode ray phosphorous.

6. Silicon Carbide

(a) It has large energy gap of 3.0eV and is regarded as high temperature semiconductor.

(b) It is extremely refractory subliming in the region of 2800°C.

(c) Because of the large band gap it is hoped that silicon carbide would be useful for very high temperature rectifier and transistors.

7. Gallium Phosphide

(a) It has band gap of 2.3eV.

(b) It can be used for LEDs which can emit green or red light.

(c) The red light is obtained with Zinc or cadmium oxide dopant.

1.32 Variation of Semiconductor conductivity, resistence and gap with Temperature

1. Conductivity

The conductivity of semiconductors increases with the increases in temperature. The conductivity (σ) of an intrinsic semiconductor depends upon the number of hole electron pairs and mobility. The number of hole-electron pairs increases with the rise in temperature, while its mobility decreases. Therefore, the conductivity of an intrinsic semiconductor increases with the increase in temperature. The conductivity at an temperature (T K) is given by

$$\sigma = \sigma_0[1 + \alpha(T - T_0)]$$

where α = the temperature coefficient.

The conductivity of extrinsic semiconductor decreases with the rise in temperature, as the number of majority carriers is almost constant and mobility decreases.

2. Resistance

Semiconductors have negative temperature coefficient of resistance, i.e., the resistance of semiconductor decreases with increase in temperature.

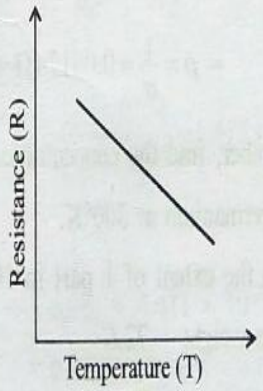


Figure 1.36 The temperature dependence of the resistance in semiconductors

3. Bond Gap

The energy gap decreases with the increase in temperature and is given by

$$E_G(T) = E_{Go} - \beta T$$

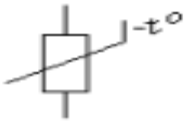
where β = a constant, whose value depends upon the nature of the material. Its value for silicon is 3.6×10^{-4} and for germanium 2.23×10^{-4}, and E_{Go} = energy gap at 0 K. Its value for silicon is 1.21 eV and for germanium, 0.785 eV.

THERMISTORS :

A thermistor is a resistance thermometer, or a resistor whose resistance is dependent on temperature. ... A PTC thermistor works a little differently. When temperature increases, the resistance increases, and when temperature decreases, resistance decreases. This type of thermistor is generally used as a fuse.

bead thermistor

disc thermistor



thermistor
circuit symbol

Thermistors are temperature sensitive semiconductors that exhibit a large change in resistance over a relatively small range of temperature. There are two main types of thermistors, positive temperature coefficient (PTC) and negative temperature coefficient (NTC). NTC thermistors exhibit the characteristic of resistance falling with increasing temperature. These are most commonly used for temperature measurement.

A *thermistor* is similar to an RTD, but a *semiconductor material* is used instead of a metal. A thermistor is a *solid state* device and has larger *sensitivity* than does an RTD. Unlike RTD's, the temperature-resistance characteristic of a thermistor is non-linear, and cannot be characterized by a single coefficient. Furthermore, unlike RTDs, the resistance of a thermistor *decreases* with increasing temperature.

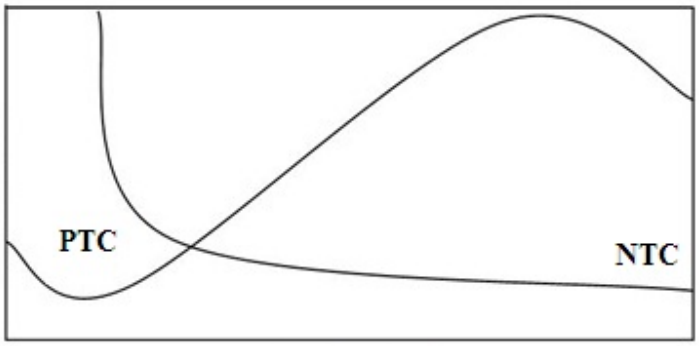
Thermistor Working Principle

Thermistors functioning is described as

- The principle obeyed by the thermistor is its dependency on-resistance values on the change in temperature.
- The value of resistance can be measured by using an ohmmeter. These are connected in series with the battery and the meter.
- The change in the resistance is depending upon the material chosen in the construction of the thermistor.
- Thermistors are considered to be a special variety of resistors. Generally, the resistor is known for limiting the amount of current in the circuit.
- But in this thermal resistors, the change in the resistance is dependent upon the change in temperature.
- If the temperature tends to increase the resistance in the circuit decreases in these special variants of resistors. It is decided based on the coefficient of temperature.

Characteristics of Thermistors

The characteristics of the thermistors changes based on whether it is of positive coefficient type or negative coefficient type. In the PTC temperature and resistance are in direct proportion whereas in NTC these are inversely related to each other.



thermistor -characteristics

Applications of Thermistor

The applications of thermistors are as follows:

1. These are compact. It can be used as a temperature sensor in Digital Thermometers.
2. In the automotive industry to measure the temperature of the coolant and the oil in the trucks as well as in cars these are preferred.
3. Household appliances make use of thermistor to increase or decrease the amount of heat required.
4. To protect the circuits from the overloading effect that is by increasing the resistance value. Hence the thermistors are considered as the circuit protection elements.
5. In the circuits of Wheat stone Bridge, Rechargeable batteries, Electronic device circuits thermistors are utilized.

UNIT-2

Review of Quantum Mechanics :

This chapter gives a brief introduction to quantum mechanics. Quantum mechanics can be thought of roughly as the study of physics on very small length scales, although there are also certain macroscopic systems it directly applies to. The descriptor "quantum" arises because in contrast with classical mechanics, certain quantities take on only discrete values. However, some quantities still take on continuous values, as we'll see. In quantum mechanics, particles have wavelike properties, and a particular wave equation, the Schrodinger equation, governs how these waves behave. The Schrodinger equation is different in a few ways from the other wave equations we've seen in this book. But these differences won't keep us from applying all of our usual strategies for solving a wave equation and dealing with the resulting solutions. In some respect, quantum mechanics is just

another example of a system governed by a wave equation. In fact, we will find below that some quantum mechanical systems have exact analogies to systems we've already studied in this book. So the results can be carried over, with no modifications whatsoever needed. However, although it is fairly straightforward to deal with the actual waves, there are many things about quantum mechanics that are a combination of subtle, perplexing, and bizarre. To name a few: the measurement problem, hidden variables along with Bell's theorem, and wave-particle duality. You'll learn all about these in an actual course on quantum mechanics. Even though there are many things that are highly confusing about quantum mechanics, the nice thing is that it's relatively easy to apply quantum mechanics to a physical system to figure out how it behaves. There is fortunately no need to understand all of the subtleties about quantum mechanics in order to use it. Of course, in most cases this isn't the best strategy to take; it's usually not a good idea to blindly forge ahead with something if you don't understand what you're actually working with. But this lack of understanding can be forgiven in the case of quantum mechanics, because no one really understands it. (Well, maybe a couple people do, but they're few and far between.) If the world waited to use quantum mechanics until it understood it, then we'd be stuck back in the 1920's. The bottom line is that quantum mechanics can be used to make predictions that are consistent with experiment. It hasn't failed us yet. So it would be foolish not to use it. Even though there are many things that are highly confusing about quantum mechanics, the nice thing is that it's relatively easy to apply quantum mechanics to a physical system to figure out how it behaves. There is fortunately no need to understand all of the subtleties about quantum mechanics in order to use it. Of course, in most cases this isn't the best strategy to take; it's usually not a good idea to blindly forge ahead with something if you don't understand what you're actually working with. But this lack of understanding can be forgiven in the case of quantum mechanics, because no one really understands it. (Well, maybe a couple people do, but they're few and far between.) If the world waited to use quantum mechanics until it understood it, then we'd be stuck back in the 1920's. The bottom line is that quantum mechanics can be used to make predictions that are consistent with experiment. It hasn't failed us yet. So it would be foolish not to use it. The main purpose of this chapter is to demonstrate how similar certain results in quantum mechanics are to earlier results we've derived in the book. You actually know a good.

A brief history Before discussing the Schrodinger wave equation, let's take a brief (and by no means comprehensive) look at the historical timeline of how quantum mechanics came about. The actual history is of course never as clean as an outline like this suggests, but we can at least get a general idea of how things

proceeded. 1900 (Planck): Max Planck proposed that light with frequency ν is emitted in quantized lumps of energy that come in integral multiples of the quantity, $E = h\nu = \bar{h}\omega$ (1) where $h \approx 6.63 \cdot 10{-}34$ J · s is Planck's constant, and $\bar{h} \equiv h/2\pi = 1.06 \cdot 10{-}34$ J · s. The frequency ν of light is generally very large (on the order of 1015 s −1 for the visible spectrum), but the smallness of h wins out, so the hν unit of energy is very small (at least on an everyday energy scale). The energy is therefore essentially continuous for most purposes. However, a puzzle in late 19th-century physics was the blackbody radiation problem. In a nutshell, the issue was that the classical (continuous) theory of light predicted that certain objects would radiate an infinite amount of energy, which of course can't be correct. Planck's hypothesis of quantized radiation not only got rid of the problem of the infinity, but also correctly predicted the shape of the power curve as a function of temperature. The results that we derived for electromagnetic waves in Chapter 8 are still true. In particular, the energy flux is given by the Poynting vector in Eq. 8.47. And E = pc for a light. Planck's hypothesis simply adds the information of how many lumps of energy a wave contains. Although strictly speaking, Planck initially thought that the quantization was only a function of the emission process and not inherent to the light itself. 1905 (Einstein): Albert Einstein stated that the quantization was in fact inherent to the light, and that the lumps can be interpreted as particles, which we now call "photons." This proposal was a result of his work on the photoelectric effect, which deals with the absorption of light and the emission of elections from a material. We know from Chapter 8 that E = pc for a light wave. (This relation also follows from Einstein's 1905 work on relativity, where he showed that E = pc for any massless particle, an example of which is a photon.) And we also know that ω = ck for a light wave. So Planck's $E = \bar{h}\omega$ relation becomes $E = \bar{h}\omega \Rightarrow pc = \bar{h}(ck) \Rightarrow p = \bar{h}k$ (2) This result relates the momentum of a photon to the wavenumber of the wave it is associated 1913 (Bohr): Niels Bohr stated that electrons in atoms have wavelike properties. This correctly explained a few things about hydrogen, in particular the quantized energy levels that were known. 1924 (de Broglie): Louis de Broglie proposed that all particles are associated with waves, where the frequency and wavenumber of the wave are given by the same relations we found above for photons, namely $E = \bar{h}\omega$ and $p = \bar{h}k$. The larger E and p are, the larger ω and k are. Even for small E and p that are typical of a photon, ω and k are very large because $\bar{h}$ is so small. So any everyday-sized particle with large (in comparison) energy and momentum values will have extremely large ω and k values. This (among other reasons) makes it virtually impossible to observe the wave nature of macroscopic amounts of matter. This proposal (that $E = \bar{h}\omega$ and $p = \bar{h}k$ also hold for massive particles) was a big step, because many things that are true for photons are not true for massive (and

nonrelativistic) particles. For example, $E = pc$ (and hence $\omega = ck$) holds only for massless particles (we'll see below how ω and k are related for massive particles). But the proposal was a reasonable one to try. And it turned out to be correct, in view of the fact that the resulting predictions agree with experiments. The fact that any particle has a wave associated with it leads to the so-called waveparticle duality. Are things particles, or waves, or both? Well, it depends what you're doing with them. Sometimes things behave like waves, sometimes they behave like particles. A vaguely true statement is that things behave like waves until a measurement takes place, at which point they behave like particles. However, approximately one million things are left unaddressed in that sentence. The wave-particle duality is one of the things that few people, if any, understand about quantum mechanics.

The Schrodinger equation

In this section we'll give a "derivation" of the Schrodinger equation. Our starting point will be the classical nonrelativistic expression for the energy of a particle, which is the sum of the kinetic and potential energies. We'll assume as usual that the potential is a function of only x. We have

In this section we'll give a "derivation" of the Schrodinger equation. Our starting point will be the classical nonrelativistic expression for the energy of a particle, which is the sum of the kinetic and potential energies. We'll assume as usual that the potential is a function of only x. We have

$$E = K + V = \frac{1}{2}mv^2 + V(x) = \frac{p^2}{2m} + V(x). \tag{3}$$

We'll now invoke de Broglie's claim that all particles can be represented as waves with frequency ω and wavenumber k, and that $E = \hbar\omega$ and $p = \hbar k$. This turns the expression for the energy into

$$\hbar\omega = \frac{\hbar^2 k^2}{2m} + V(x). \tag{4}$$

A wave with frequency ω and wavenumber k can be written as usual as $\psi(x,t) = Ae^{i(kx-\omega t)}$ (the convention is to put a minus sign in front of the ωt). In 3-D we would have $\psi(\mathbf{r},t) = Ae^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)}$, but let's just deal with 1-D. We now note that

$$\begin{aligned} \frac{\partial\psi}{\partial t} &= -i\omega\psi \quad\Longrightarrow\quad \omega\psi = i\frac{\partial\psi}{\partial t}, \qquad \text{and} \\ \frac{\partial^2\psi}{\partial x^2} &= -k^2\psi \quad\Longrightarrow\quad k^2\psi = -\frac{\partial^2\psi}{\partial x^2}. \end{aligned} \tag{5}$$

If we multiply the energy equation in Eq. (4) by ψ, and then plug in these relations, we obtain

$$\hbar(\omega\psi) = \frac{\hbar^2}{2m}(k^2\psi) + V(x)\psi \quad\Longrightarrow\quad \boxed{i\hbar\frac{\partial\psi}{\partial t} = \frac{-\hbar^2}{2m}\cdot\frac{\partial^2\psi}{\partial x^2} + V\psi} \tag{6}$$

This is the *time-dependent Schrodinger equation*. If we put the x and t arguments back in, the equation takes the form,

$$i\hbar\frac{\partial\psi(x,t)}{\partial t} = \frac{-\hbar^2}{2m}\cdot\frac{\partial^2\psi(x,t)}{\partial x^2} + V(x)\psi(x,t). \tag{7}$$

In 3-D, the x dependence turns into dependence on all three coordinates (x, y, z), and the ∂ 2ψ/∂x2 term becomes ∇2ψ (the sum of the second derivatives). Remember that Born's (correct) interpretation of ψ(x) is that |ψ(x)| 2 gives the probability of finding the particle at position x. Having successfully produced the time-dependent Schrodinger equation, we should ask: Did the above reasoning actually prove that the Schrodinger equation is valid? No, it didn't, for three reasons. 1. The reasoning is based on de Broglie's assumption that there is a wave associated with every particle, and also on the assumption that the ω and k of the wave are related to E and p via Planck's constant in Eqs. (1) and (2). We had to accept these assumptions on faith. 2. Said in a different way, it is impossible to actually prove anything in physics. All we can do is make an educated guess at a theory, and then do experiments to try to show

The Schrodinger equation actually isn't valid, so there's certainly no way that we proved it. Consistent with the above point concerning limiting cases, the quantum theory based on Schrodinger's equation is just a limiting theory of a more correct one, which happens to be quantum field theory (which unifies quantum mechanics with special relativity). This is turn must be a limiting theory of yet another more correct one, because it doesn't incorporate gravity. Eventually there will be one theory that covers everything (although this point can be debated), but we're definitely not there yet. Due to the "i" that appears in Eq. (6), $\psi(x)$ is complex. And in contrast with waves in classical mechanics, the entire complex function now matters in quantum mechanics. We won't be taking the real part in the end. Up to this point in the book, the use of complex functions was simply a matter of convenience, because it is easier to work with exponentials than trig functions. Only the real part mattered (or imaginary part – take your pick, but not both). But in quantum mechanics the whole complex wavefunction is relevant. However, the theory is structured in such a way that anything you might want to measure (position, momentum, energy, etc.) will always turn out to be a real quantity. This is a necessary feature of any valid theory, of course, because you're not going to go out and measure a distance of $2 + 5i$ meters, or pay an electrical bill of $17 + 6i$ kilowatt hours. As mentioned in the introduction to this chapter, there is an endless number of difficult questions about quantum mechanics that can be discussed. But in this short introduction to the subject, let's just accept Schrodinger's equation as valid, and see where it takes us.

string can be the superposition of various normal modes with definite ω's. The same reasoning applies here as with all the other waves we've discussed: From Fourier analysis and from the linearity of the Schrodinger equation, we can build up any general wavefunction from ones with specific energies. Because of this, it suffices to consider the time-independent Schrodinger equation. The solutions for that equation form a basis for all possible solutions.[1]

Continuing with our standard strategy of guessing exponentials, we'll let $\psi(x) = Ae^{ikx}$. Plugging this into Eq. (9) and canceling the e^{ikx} gives (going back to the $\hbar\omega$ instead of E)

$$\hbar\omega = -\frac{\hbar^2}{2m}(-k^2) + V(x) \quad \Longrightarrow \quad \hbar\omega = \frac{\hbar^2 k^2}{2m} + V(x). \tag{10}$$

This is simply Eq. (4), so we've ended up back where we started, as expected. However, our goal here was to show how the Schrodinger equation can be solved from scratch, without knowing where it came from.

Eq. (10) is (sort of) a dispersion relation. If $V(x)$ is a constant C in a given region, then the relation between ω and k (namely $\omega = \hbar k^2/2m + C$) is independent of x, so we have a nice sinusoidal wavefunction (or exponential, if k is imaginary). However, if $V(x)$ isn't constant, then the wavefunction isn't characterized by a unique wavenumber. So a function of the form e^{ikx} doesn't work as a solution for $\psi(x)$. (A Fourier superposition can certainly work, since any function can be expressed that way, but a single e^{ikx} by itself doesn't work.) This is similar to the case where the density of a string isn't constant. We don't obtain sinusoidal waves there either.

string can be the superposition of various normal modes with definite ω's. The same reasoning applies here as with all the other waves we've discussed: From Fourier analysis and from the linearity of the Schrodinger equation, we can build up any general wavefunction from ones with specific energies. Because of this, it suffices to consider the time-independent Schrodinger equation. The solutions for that equation form a basis for all possible solutions.[1]

Continuing with our standard strategy of guessing exponentials, we'll let $\psi(x) = Ae^{ikx}$. Plugging this into Eq. (9) and canceling the e^{ikx} gives (going back to the $\hbar\omega$ instead of E)

$$\hbar\omega = -\frac{\hbar^2}{2m}(-k^2) + V(x) \quad \Longrightarrow \quad \hbar\omega = \frac{\hbar^2 k^2}{2m} + V(x). \tag{10}$$

This is simply Eq. (4), so we've ended up back where we started, as expected. However, our goal here was to show how the Schrodinger equation can be solved from scratch, without knowing where it came from.

Eq. (10) is (sort of) a dispersion relation. If $V(x)$ is a constant C in a given region, then the relation between ω and k (namely $\omega = \hbar k^2/2m + C$) is independent of x, so we have a nice sinusoidal wavefunction (or exponential, if k is imaginary). However, if $V(x)$ isn't constant, then the wavefunction isn't characterized by a unique wavenumber. So a function of the form e^{ikx} doesn't work as a solution for $\psi(x)$. (A Fourier superposition can certainly work, since any function can be expressed that way, but a single e^{ikx} by itself doesn't work.) This is similar to the case where the density of a string isn't constant. We don't obtain sinusoidal waves there either.

string can be the superposition of various normal modes with definite ω's. The same reasoning applies here as with all the other waves we've discussed: From Fourier analysis and from the linearity of the Schrodinger equation, we can build up any general wavefunction from ones with specific energies. Because of this, it suffices to consider the time-independent Schrodinger equation. The solutions for that equation form a basis for all possible solutions.[1]

Continuing with our standard strategy of guessing exponentials, we'll let $\psi(x) = Ae^{ikx}$. Plugging this into Eq. (9) and canceling the e^{ikx} gives (going back to the $\hbar\omega$ instead of E)

$$\hbar\omega = -\frac{\hbar^2}{2m}(-k^2) + V(x) \quad \Longrightarrow \quad \hbar\omega = \frac{\hbar^2 k^2}{2m} + V(x). \tag{10}$$

This is simply Eq. (4), so we've ended up back where we started, as expected. However, our goal here was to show how the Schrodinger equation can be solved from scratch, without knowing where it came from.

Eq. (10) is (sort of) a dispersion relation. If $V(x)$ is a constant C in a given region, then the relation between ω and k (namely $\omega = \hbar k^2/2m + C$) is independent of x, so we have a nice sinusoidal wavefunction (or exponential, if k is imaginary). However, if $V(x)$ isn't constant, then the wavefunction isn't characterized by a unique wavenumber. So a function of the form e^{ikx} doesn't work as a solution for $\psi(x)$. (A Fourier superposition can certainly work, since any function can be expressed that way, but a single e^{ikx} by itself doesn't work.) This is similar to the case where the density of a string isn't constant. We don't obtain sinusoidal waves there either.

f

Solving the equation

If we put aside the profound implications of the Schrodinger equation and regard it as simply a mathematical equation, then it's just another wave equation. We already know the solution, of course, because we used the function $\psi(x,t) = Ae^{i(kx-\omega t)}$ to produce Eqs. (5) and (6) in the first place. But let's pretend that we don't know this, and let's solve the Schrodinger equation as if we were given it out of the blue.

As always, we'll guess an exponential solution. If we first look at exponential behavior in the time coordinate, our guess is $\psi(x,t) = e^{-i\omega t}f(x)$ (the minus sign here is convention). Plugging this into Eq. (7) and canceling the $e^{-i\omega t}$ yields

$$\hbar\omega f(x) = -\frac{\hbar^2}{2m}\frac{\partial^2 f(x)}{\partial x^2} + V(x)f(x). \tag{8}$$

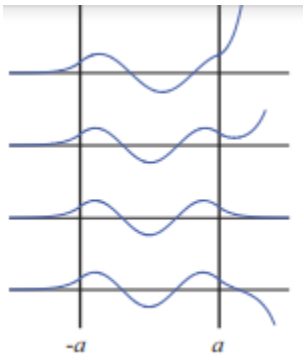
But from Eq. (1), we have $\hbar\omega = E$. And we'll now replace $f(x)$ with $\psi(x)$. This might cause a little confusion, since we've already used ψ to denote the entire wavefunction $\psi(x,t)$. However, it is general convention to also use the letter ψ to denote the spatial part. So we now have

$$\boxed{E\,\psi(x) = -\frac{\hbar^2}{2m}\frac{\partial^2\psi(x)}{\partial x^2} + V(x)\psi(x)} \tag{9}$$

This is called the *time-independent Schrodinger equation*. This equation is more restrictive than the original time-*dependent* Schrodinger equation, because it assumes that the particle/wave has a definite energy (that is, a definite ω). In general, a particle can be in a state that is the superposition of states with various definite energies, just like the motion of a

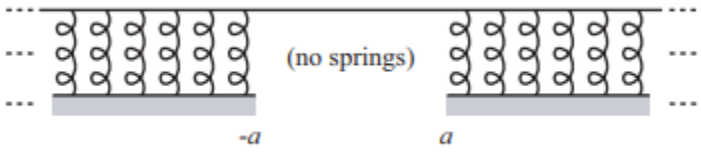
Explanation of the quantized energy levels

The most important thing to note about these states is that they are discrete. In the infinitewell case, this discreteness was clear because an integral number of half wavelengths needed to fit into the well (because $\psi = 0$ at the boundaries). The discreteness isn't so obvious in the finite-well case (because ψ 6= 0 at the boundaries), but it is still reasonably easy to see. There are two ways to understand it. First, the not-so-enlightening way is to note that we initially had 7 equation and 7 unknowns. So all the unknowns, including E, are determined. There may be different discrete solutions, but at least we know that we can't just pick a random value for E and expect it to work. Second, the more physical and enlightening way is the following. If we choose a random value of E, it probably won't yield an allowable function ψ, and here's why. Pick an arbitrary value of the coefficient A1 in Eq. (20), and set B1 = 0. So we have an exponentially decaying function in the left region, which behaves properly at $x = -\infty$. Since E determines κ, we know everything about the $A1e\ \kappa x$ function. Now pick a point x0 in the left region, and imagine marching rightward on the x axis. We claim that all of the subsequent values of ψ are completely determined, all the way up to $x = +\infty$. This is clear in the left region, because we know what the function is. But it is also true in the middle and right regions, because we know the values of ψ, ψ 0 , and ψ 00 at any given point, so we can recursively find these three values at the "next" point as we march along. More precisely: (a) from the definition of the derivative, the values of ψ and ψ 0 at a given point yield the value of ψ at the next point, (b) likewise, the values of ψ 0 and ψ 00 at a given point yield the value of ψ 0 at the next point, and (c) finally, the value of ψ at a given point yields the value of ψ 00 at that point, via the Schrodinger equation, Eq. (9). So we can recursively march all the way up to $x = +\infty$, and the entire function is determined. There is no freedom whatsoever in what the function turns out to be. A particular choice of E and A1 might yield the first (top) function shown in Fig. 7. It has -a a Figure 7 the correct exponential and oscillatory nature in the left and middle regions, respectively. But in the right region it apparently has an exponentially growing piece. Because it diverges at $x = +\infty$, this function isn't an allowable one. So it is impossible for the energy to take on the value that we chose. We can try to remedy the divergence at $x = +\infty$ by increasing the value of E. This will make ψ oscillate quicker inside the well, so that it encounters the x = a boundary with a negative slope. We then might end up with the second function shown in Fig. 7. We still have the divergence at $x = +\infty$, so again we have an invalid ψ. If we increase E a little more, by precisely the right amount, then we'll end up with the third function shown in Fig.



As with the infinite square well, the finite square well also has a direct analogy with a setup involving a string. Recall the discussion of the string/spring system in Section 6.2.2, involving evanescent waves and a low-frequency cutoff. Consider the system shown in Fig. 9, where the springs extend to $x = \pm\infty$. From Section 6.2.2, we know that if the frequency is below the cutoff frequency, then we have a sinusoidal wave in the middle region, but an evanescent wave (that is, an exponentially decaying wave) in the side regions. This is exactly what we have in the quantum-mechanical finite-well setup. In both setups the most general forms of the waves (as functions of x) in the different regions are given by Eq. (20) (but we need to take the real part in the string/spring case). And the boundary conditions are the same: continuity of ψ and ψ 0 at the boundaries, and $\psi = 0$ at $x = \pm\infty$. If you've solved for one $\psi(x)$, you've solved for the other.

TUNNELING :



Consider the string/spring system that is the "opposite" of the system shown in Fig. 9. So we now have springs in the middle region, and a normal string in the outer regions, as shown in Fig. 10.

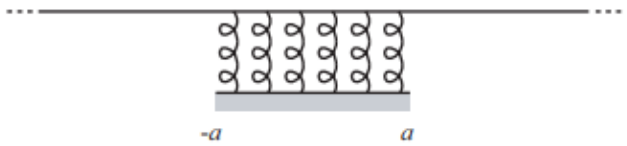


Figure 10

If a rightward-traveling wave comes in from the left, and if the frequency ω is less than the cutoff frequency ω_s, then we will have an evanescent wave in the middle region. And we will also have rightward-traveling wave in the right region (but no leftward-traveling wave, because we're not sending anything in from the right). So the waves in the three regions take the general form,

$$\begin{aligned} x < -a: &\quad \psi_1(x) = A_1 e^{ikx} + B_1 e^{-ikx}, \\ -a < x < a: &\quad \psi_2(x) = A_2 e^{\kappa x} + B_2 e^{-\kappa x}, \\ x > a: &\quad \psi_3(x) = A_3 e^{ikx}. \end{aligned} \tag{24}$$

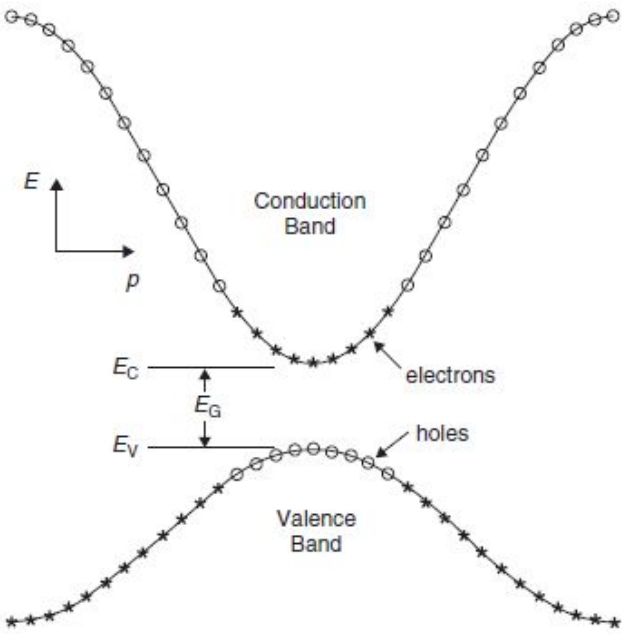
We've given only the x dependence here. All of these waves are multiplied by the same function of t, namely $e^{-i\omega t}$. The ratio B_1/A_1 is the reflection coefficient, and the ratio A_3/A_1

E-k diagram

An **E-k diagram** shows characteristics of a particular semiconductor material. It shows the relationship between the energy and momentum of available quantum mechanical states for electrons in the material.The effective mass of electrons and holes in the material. This is given by the curvature of each of the bands.

An E-k diagram shows characteristics of a particular semiconductor material. It shows the relationship between the energy and momentum of available quantum mechanical states for electrons in the material.

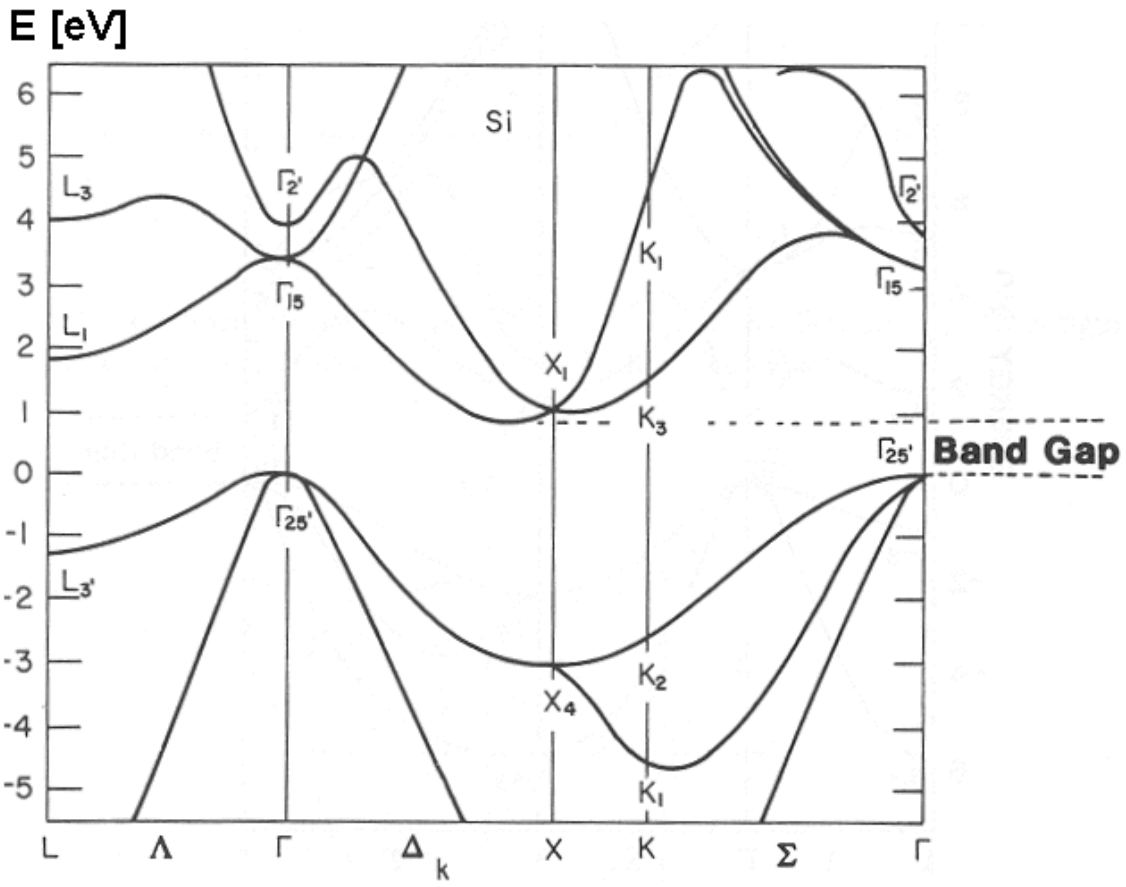
First, consider a basic E-k band diagram like this one (the x-axis can be either momentum, p, or wavenumber, k, since $p = \hbar k$):



In this diagram you can see a few things:

- The band gap (E_G), which is the difference in energy between the top of the valence band and the bottom of the conduction band.
- The effective mass of electrons and holes in the material. This is given by the curvature of each of the bands.
- This diagram indicates (diagramatically) how the actual electron states are equally spaced in k-space. Which means that the density of states in E ($\rho(E)$) depends on the slope of the E-k curve.

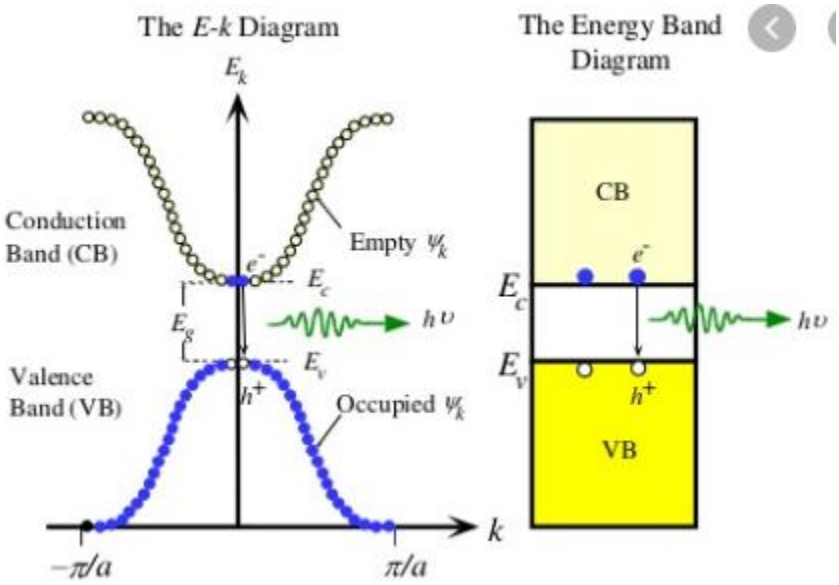
There is a more complex form of E-k diagram that shows the relationship for different directions of k relative to the crystal lattice:



Here, the greek letters (Γ, Δ, K, etc.) on the x-axis indicate different directions of k relative to the crystal axes.

In addition to showing the effective mass at different band extrema, this also shows that the effective mass varies depending on the direction of conduction relative to the crystal orientation.

This type diagram also shows whether the material is a direct-gap or indirect-gap semiconductor. Direct gap is when the valence band maximum and conduction band minimum occur at the same location in k-space. This is important in optoelectronics because only direct gap materials (like GaAs, but not including silicon) have efficient radiative absorption and emission, which is what makes LEDs and laser diodes work.



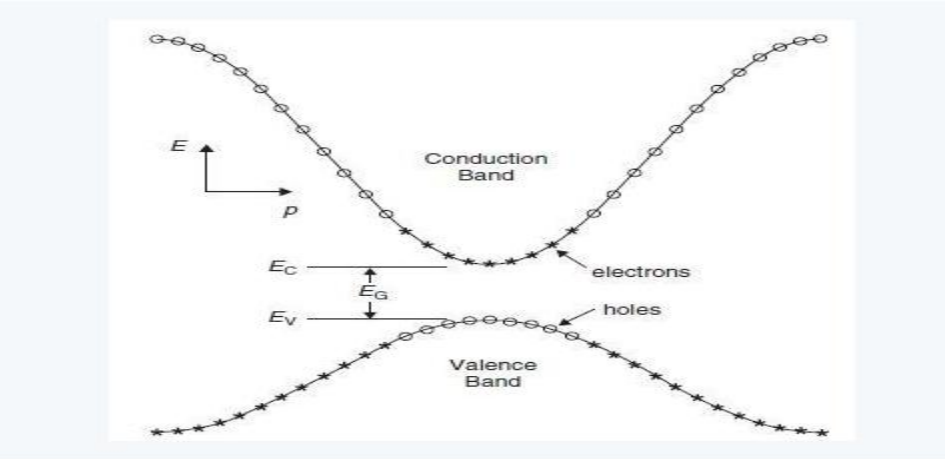
An E-k diagram shows characteristics of a particular semiconductor material. It shows the relationship between the energy and momentum of available quantum mechanical states for electrons in the material. First, consider a basic E-k band diagram like this one (the x-axis can be either momentum, pp, or wavenumber, kk, since p=ħkp=ħk):

In this diagram you can see a few things:

The band gap (EG), which is the difference in energy between the top of the valence band and the bottom of the conduction band.

The effective mass of electrons and holes in the material. This is given by the curvature of each of the bands.

This diagram indicates (diagramatically) how the actual electron states are equally spaced in k-space. Which means that the density of states in E (ρ(E)ρ(E)) depends on the slope of the E-k curve.



DRIFT AND DIFFUSION CURRENT

Drift current is electric **current** due to the motion of charge carriers under the influence of an external electric field while **diffusion current** is electric **current** due to the **diffusion** of carriers leading to a change in carrier concentration.

Diffusion current	**Drift current**
(WHEN). Diffusion current occurs even though there isn't an electric field applied to the semiconductor .	Drift current depends on the electric field applied on the p-n junction diode.
(WHY) It depends on constants Dp and Dn, and +q and -q, for holes and electrons respectively but it is independent of permittivity.	It depends upon permittivity.
(WHERE) Direction of the diffusion current depends on the change in the carrier concentrations.	Direction of the drift current depends on the polarity of the applied field.

In a semiconductor, the majority and minority charge carriers will exit in p-type or n-type. Because both the types of semiconductors will present over a single crystal at the center so that PN-junction can be formed. When the doping of this junction diode is done non-uniformly then charge carriers movement will be an exit from high to low concentration which leads to the recombination of carriers as well as to the diffusion process. There is an additional method is also occurs based on the applied electric field namely drift current. This article discusses the main differences between drift current and diffusion current.

In a semiconductor material, the drift, as well as diffusion currents, will occur. Semiconductors are fabricated with two kinds of materials namely p-type as well as n-type. There are several kinds of switching devices available in the market like transistors, diodes, etc. These are designed by placing one material among other materials so that the material's conducting property can be modified.

$$J = qn\mu E + qD\frac{dn}{dx} \quad (1)$$

here D is the diffusion coefficient for the electron in the considered medium, n is the number of electrons per unit volume (i.e. number density), q is the magnitude of charge of an lectron, μ is electron mobility in the medium, and $E = -d\Phi/dx$ (Φ potential difference) is the electric field as the potential gradient of the electric potential. According to the Einstein elation on electrical mobility $D = \mu V_t$ and $V_t = kT/q$. Thus, substituting E for the potential gradient in the above equation (1) and multiplying both sides with $\exp(-\Phi/V_t)$, (1) ecomes:

$$Je^{-\Phi/V_t} = qD\left(\frac{-n}{V_t} * \frac{d\Phi}{dx} + \frac{dn}{dx}\right)e^{-\Phi/V_t} = qD\frac{d}{dx}(ne^{-\Phi/V_t}) \quad (2)$$

ntegrating equation (2) over the depletion region gives

$$J = \frac{qDne^{-\Phi/V_t}\Big|_0^{x_d}}{\int_0^{x_d} e^{-\Phi/V_t}dx}$$

hich can be written as

$$J = \frac{qDN_c e^{-\Phi_B/V_t}\left[e^{V_a/V_t} - 1\right]}{\int_0^{x_d} e^{-\Phi^*/V_t}dx} \quad (3)$$

here

$$\Phi^* = \Phi_B + \Phi_i - V_a$$

he denominator in equation (3) can be solved by using the following equation:

$$\Phi = -\frac{qN_d}{2E_s(x - x_d)^2}$$

$$\Phi^* = \frac{qN_d x}{E_s}\left(x_d - \frac{x}{2}\right) = (\Phi_i - V_a)\frac{x}{x_d} \tag{4}$$

Since the $x << x_d$ the term $(x_d - x/2) \approx x_d$, using this approximation equation (3) is solved as follows:

$$\int_0^{x_d} e^{-\Phi^*/V_t} dx = x_d \frac{\Phi_i - V_a}{V_t},$$

since $(\Phi_i - V_a) > V_t$. One obtains the equation of current caused due to diffusion:

$$J = \frac{q_2 D N_c}{V_t}\left[\frac{2q}{E_s}(\Phi_i - V_a)N_d\right]^{1/2} e^{-\Phi_B/V_t}\left(e^{V_a/V_t} - 1\right) \tag{5}$$

From equation (5), one can observe that the current depends exponentially on the input voltage V_a, also the barrier height Φ_B. From equation (5), V_a can be written as the function of electric field intensity, which is as follows:

$$E_{\max} = \left[\frac{2q}{E_s}(\Phi_i - V_a)N_d\right]^{1/2} \tag{6}$$

Substituting equation (6) in equation (5) gives:

$$J = q\mu E_{\max} N_c e^{-\Phi_B/V_t}\left(e^{V_a/V_t} - 1\right) \tag{7}$$

From equation (7), one can observe that when a zero voltage is applied to the semi-conductor diode, the drift current totally balances the diffusion current. Hence, the net current in a semiconductor diode at zero potential is always zero.

Mobility and Conductivity :-

fast they are

$$V_d \propto E$$

$$\boxed{V_d = \mu E}$$

where μ is mobility of the electron in $m^2/Volt\text{-}second$.

$$\mu = \frac{\text{drift velocity}}{\text{Field Intensity}} = \frac{V}{E}$$

$$= \frac{m/sec}{Volt/m} = m^2/Volt\text{-}sec.$$

Mobility is defined as the average partical drift velocity per unit electric field.

Mobility mean how quickly charge carriers move from one place to another place through out the crystal.

n per cubic meter, ... density J is

$$J = nqV_d$$

$\therefore V_d = \mu E$

$$J = nq\mu E$$

$\therefore nq\mu = \sigma$

$$J = \sigma E$$

where (σ) is the conductivity of the conductor in $(ohm-meter)^{-1}$. or mho/m

or S/m (siemens per meter)

Conductivity (Conductance):-

It is the reciprocal of electrical resistivity and measures a material's ability to conduct an electric current. It is commonly represented by the Greek letters σ (sigma). It's SI unit is siemens per meter (S/m).

* conductivity denotes current carrying capacity of material

Conductivity and Mobility for Intrinsic Semiconductor:-

Current density in metal is,

$$J = qn\mu E$$

The current density, due to the movement of electrons, is given by,

$$J_n = q \cdot n \cdot \mu_n \cdot E \quad \text{---(1)}$$

Where:-

q = charg on an electron

n = Electron concentration in intrinsic semiconductor or no. of electron in semiconductor

μ_n = Mobility of electron in semiconductor

E = Electric field applied.

Similarly, the current density due to the movement of holes in an intrinsic semiconductor,

$$J_p = q \cdot p \cdot \mu_p \cdot E$$

where:- q = charge on a hole

p = Hole concentration

μ_p = Mobility of holes

E = Electric field.

Total Density:-

$$J = J_n + J_p$$

$$J = q \cdot n \cdot \mu_n E + q \cdot p \cdot \mu_p E$$

$$J = q(n\mu_n + p\mu_p)E$$

Conductivity of N-type and P-type Semiconductor:-

- For N-type :- $n >> p$

$$\sigma = q \cdot n \cdot \mu_n$$

- For P-type :- $p >> n$

$$\sigma = q \cdot p \cdot \mu_p$$

- For Intrinsic semiconductor :-

$$\sigma_i = q n_i (\mu_n + \mu_p)$$

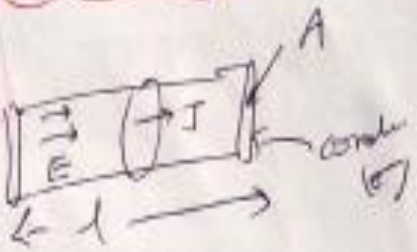
$n = p = n_i$,

$$\sigma_i = q [n_i \mu_n + n_i \mu_p]$$

$$\sigma_i = q [n \mu_n + p \mu_p]$$

Conductivity:-

Charge Densities in Semiconductor :-



$J = I/A$

$\vec{J} = \sigma \vec{E}$

$\boxed{J = \sigma E}$

$V = IR$ — (1)

$\frac{V}{l} = E$

$\Rightarrow V = El$ — (2)

$IR = El$

$I \cdot \rho \frac{l}{A} = El$ $\quad \therefore R = \rho \frac{l}{A}$

$\Rightarrow j \cdot \rho = E$

$\Rightarrow j = \frac{1}{\rho} E$

$\boxed{J = \sigma E}$ $\quad \therefore \sigma = \frac{1}{\rho}$

Resistor

A resistor is a two terminal electrical or electronic component that resist an electric current by producing a voltage ~~drop~~ drop between its terminals in accordance with ohm's Law.

$$R = V/I$$

Types →

① **General purpose :-** - General purpose resistor are of
- Carbon composition.
- They are cheap
- They are not so stable and their tolerance is 5 to 20% in value

② **High stability :-** High stability resistors include metal-film, cermet metal foil and wire wound resistor. their tolerance are low and range from 0.1 to 5%.

While selecting a resistor, the following points are to be considered:

1. Value, size, shape leads
2. Stability
3. Tolerance
4. Temp coefficient of resistance
5. Power Ratings
6. Max. operating voltage
7. Reliability
8. Change in resistance value with freq.
9. Load life environmental condition any age.

* In an ideal resistor, the resistance remains constant regardless.

Characteristic of Resistor:-

(1) Temperature Coefficient:-

Resistance changes with change in temp^r. The temp^r coefficient of resistance is defined as the percentage change in resistance per unit change in temp^r. The temp^r. coeff. can be positive or negative. depending upon whether the resistance is increases or decreases with the rise of temp^r.

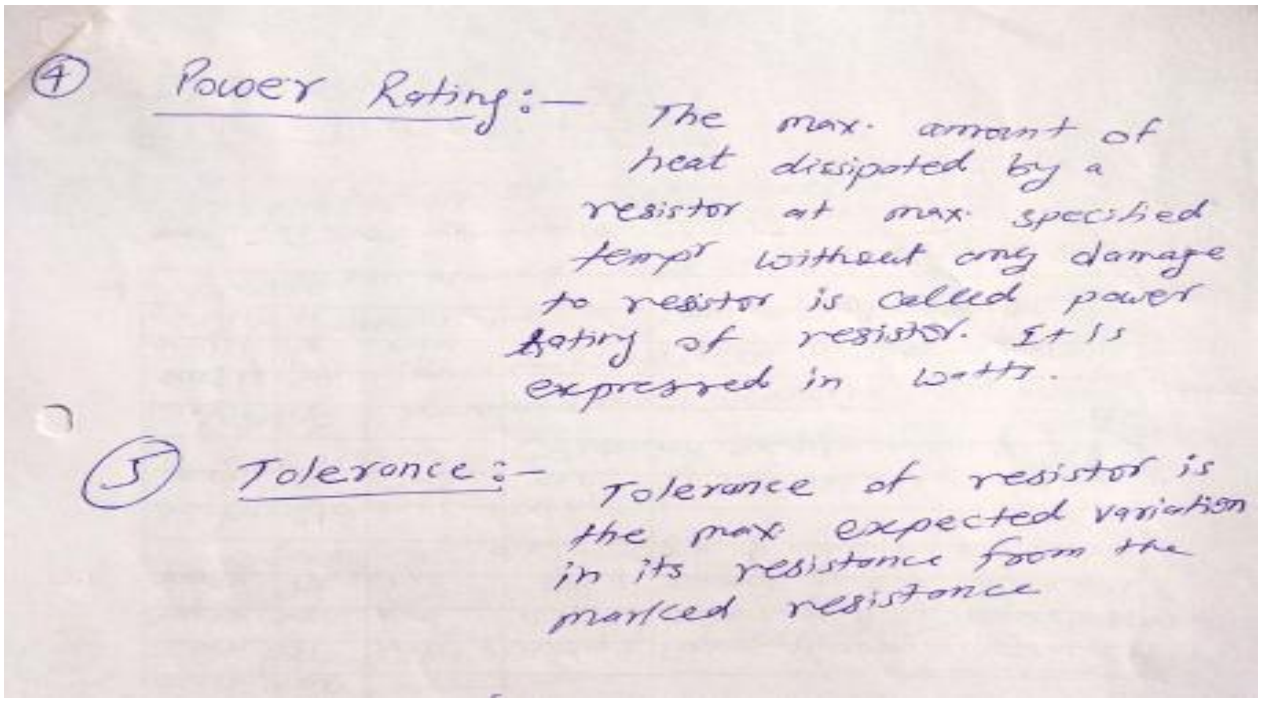
→ The temp^r. coeffi- is expressed in parts per million per degree centigrade (ppm/°C)

(2) Voltage Coefficient of Resistance:-

The voltage coefficient of resistance is defined as the percentage change in the value of resistance per unit applied voltage. → (ppm/V) => expressed

(3) Maximum Voltage Rating:-

The max. voltage that can be applied to a resistor without any damage is called max. voltage rating.



④ Power Rating :— The max. amount of heat dissipated by a resistor at max. specified temp without any damage to resistor is called power rating of resistor. It is expressed in watt.

⑤ Tolerance :— Tolerance of resistor is the max expected variation in its resistance from the marked resistance

Design of Resistors :-

(1) Preparation of Substrate:-

→ Substrate should hold the resistive element. It should be a good ~~insu~~ insulator. The overall characteristics of resistors are highly dependant upon that of substrate. Ceramic such as stellite, porcelain, mullite, forsterite and alumina are used as substrate

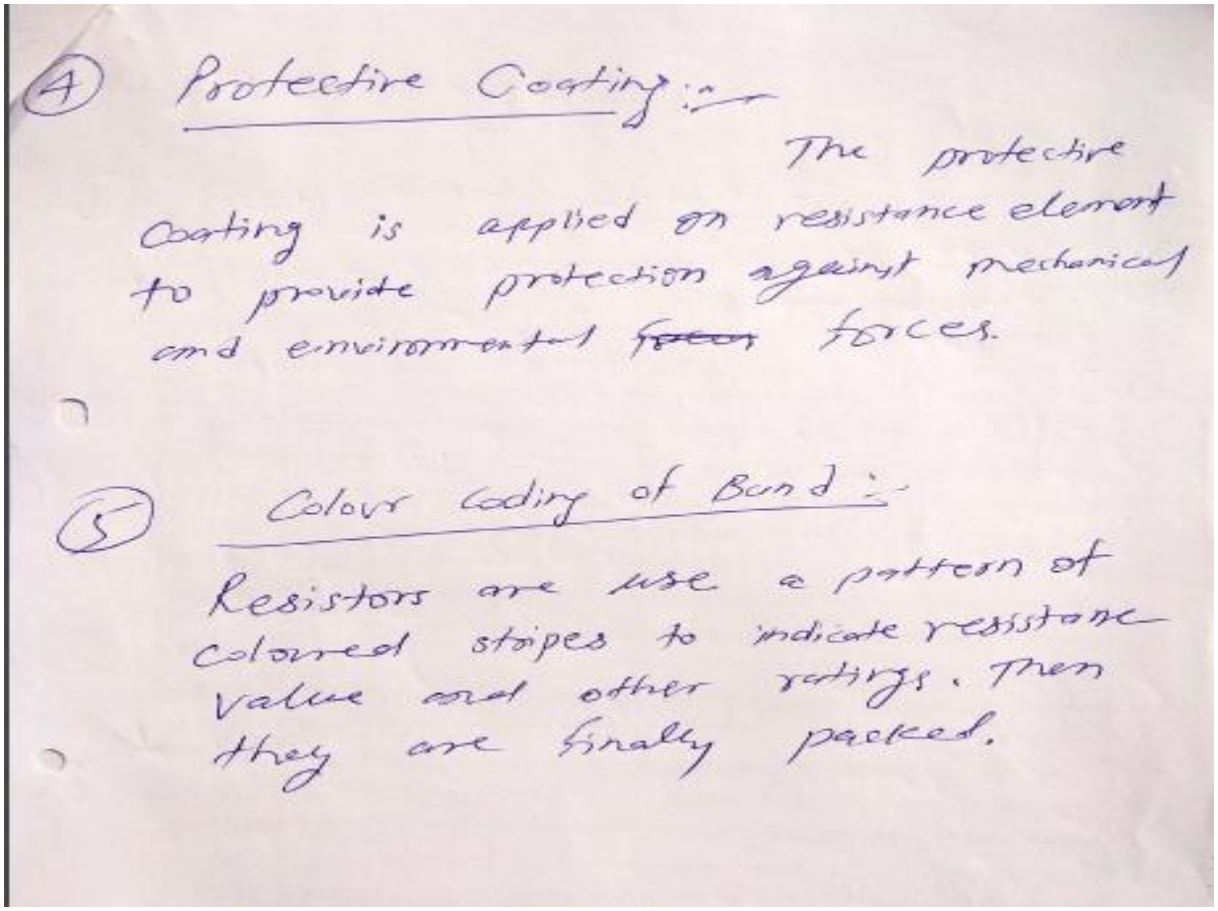
(2) - Preparation of Resistive Element:-

Generally used resistive material are metals alloys, carbon, graphite with binders

→ Resistive elements may be in the form of film or slug.

(3) - End connections or Terminal fittings:-

Terminals are used to make electrical connection of the resistive element to the circuit



(A) Protective Coating :-

The protective coating is applied on resistance element to provide protection against mechanical and environmental forces.

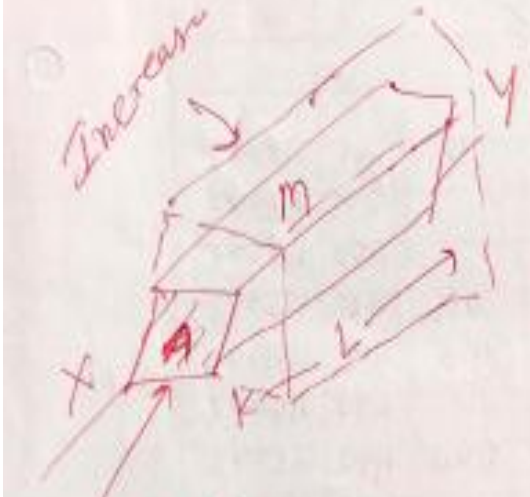
(B) Colour Coding of Band :-

Resistors are use a pattern of colored stripes to indicate resistance value and other ratings. Then they are finally packed.

Resistor, Resistance, sheet Resistance:

$R \propto L$ ✓

$R \propto \frac{1}{A}$ ✓



$R = \rho \frac{L}{A}$

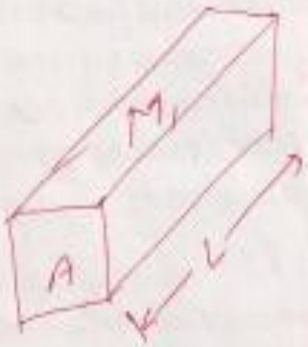
ρ = Resistivity

Cross section area

$\rightarrow \rho = \frac{RA}{L}$

$\rightarrow \rho = \frac{\Omega m^2}{m} \Rightarrow \Omega\text{-}m$ (Unit of Resistivity)

(1)



$R_1 = \rho_1 \frac{L}{A}$

ρ_1 → resistivity of material (M_1)

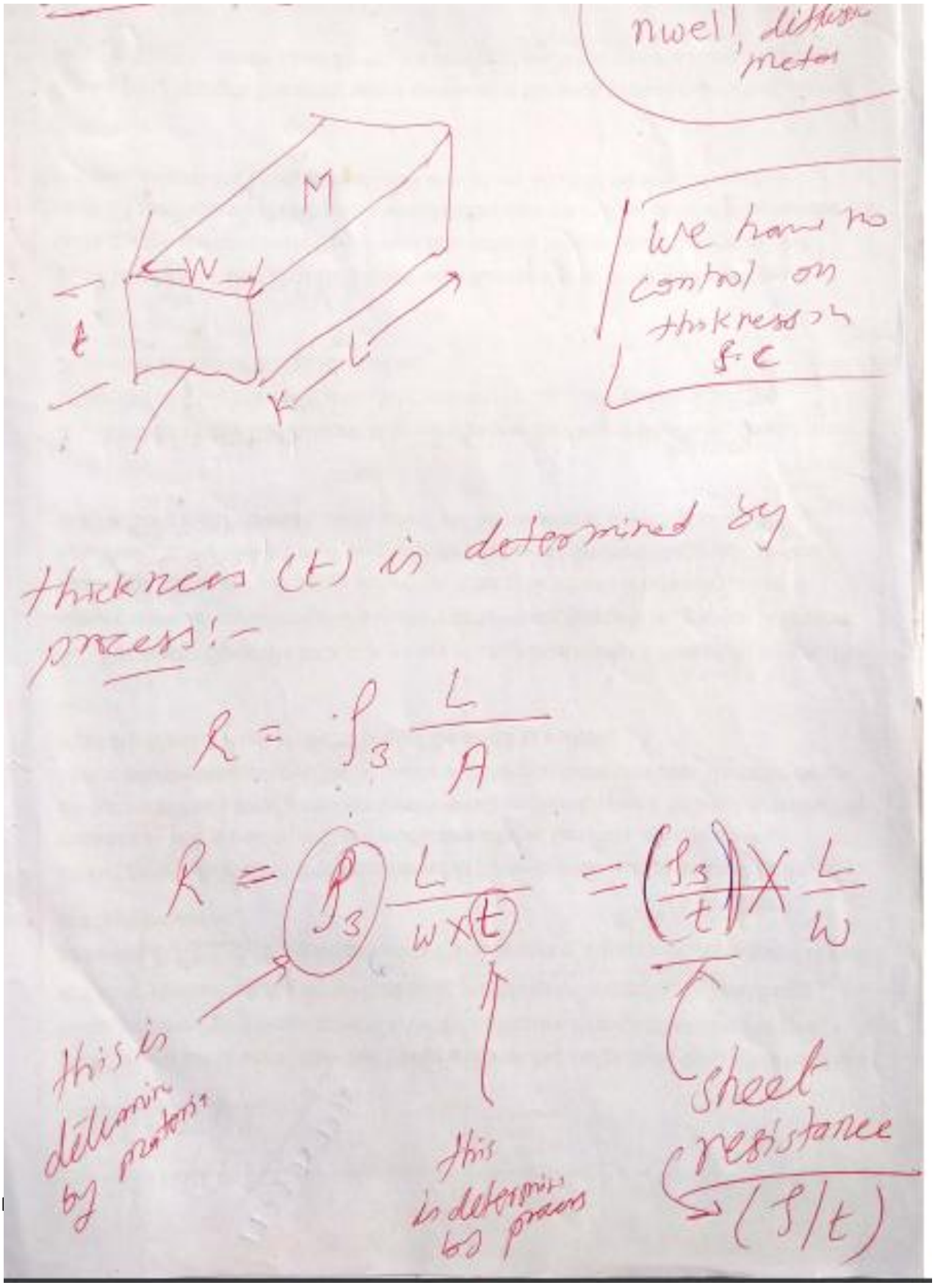
$R_3 = \rho_1 \frac{L}{A}$

Same to same material

$R_1 = R_3$

$R_2 = \rho_2 \frac{L}{A}$

ρ_2 → resistivity of material (M_2)

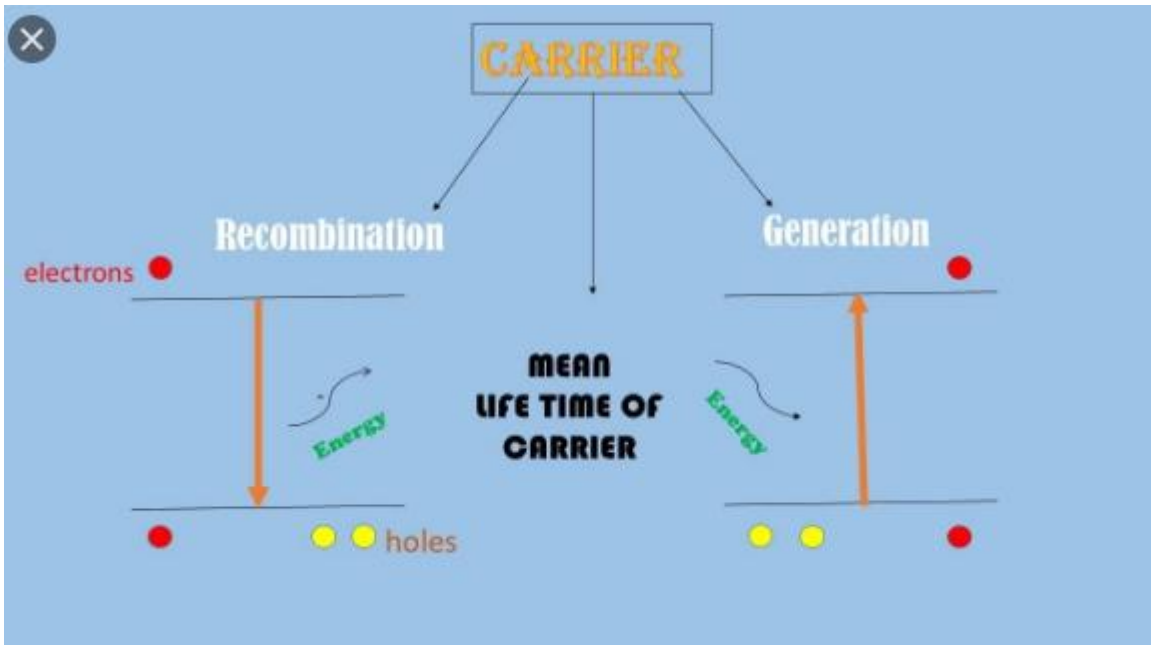

well, diffusion
metal
M
W
t
L
We have no control on thickness of S.C
Thickness (t) is determined by process!-
$R = \rho_s \frac{L}{A}$
$R = \rho_s \frac{L}{W \times t} = \left(\frac{\rho_s}{t}\right) \times \frac{L}{W}$
This is determine by process
This is determine by process
Sheet resistance
(ρ_s/t)

UNIT -3

GENERATION AND RECOMBINATION OF CARRIERS:

Carrier generation describes processes by which electrons gain energy and move from the valence band to the conduction band, producing two mobile **carriers**. while **recombination** describes processes by which a conduction band electron loses energy and re-occupies the energy state of an electron hole in the valence band. In the solid-state physics of semiconductors, **carrier generation** and **carrier recombination** are processes by which mobile charge carriers (electrons and electron holes) are created and eliminated. Carrier generation and recombination processes are fundamental to the operation of many optoelectronic semiconductor devices, such as photodiodes, light-emitting diodes and laser diodes. They are also critical to a full analysis of p-n junction devices such as bipolar junction transistors and p-n junction diodes. The **electron–hole pair** is the fundamental unit of generation and recombination in inorganic semiconductors, corresponding to an electron transitioning between the valence band and the conduction band where generation of electron is a transition from the valence band to the conduction band and recombination leads to a reverse transition.

Carrier generation describes processes by which electrons gain energy and move from the valence band to the conduction band, producing two mobile **carriers**; while **recombination** describes processes by which a conduction band electron loses energy and re-occupies the energy state of an electron hole in the valence band.


CARRIER
Recombination
Generation
electrons
MEAN
LIFE TIME OF
CARRIER
Energy
Energy
holes

Generation and Recombination :-

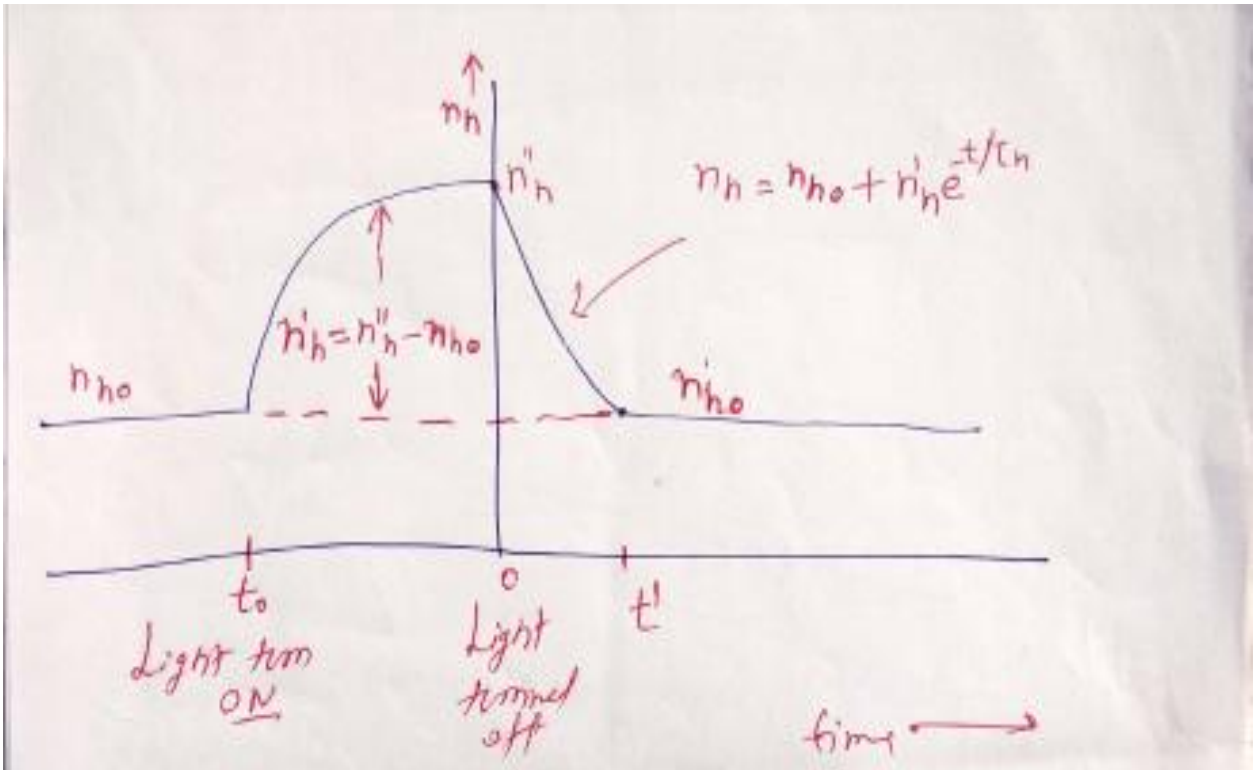
→ In an Intrinsic SC the concentration of electrons and holes are always equal.

→ Due to thermal energy, new electron-hole pairs are created. While at the same time other electron-hole pairs disappear due to recombination.

→ before recombination the electron-hole pair exist for some time. this time is called mean life time. It vary in the range of 10^{-4} sec to 10^{-9} sec.

τ_h - Life time of hole
τ_n - Life time of e^-

→ We consider a N-type Ge bar
→ Let the concentration of holes and e^- be n_{h_0} and n_{e_0} respectively
→ Under steady-state, rate of generation becomes equal to rate of recombination.



(Variation of hole concentration with time in an N-type Ge bar)

$t = t_0 \rightarrow$ light falls on germanium bar, it creats excess electron and hole pairs. After some time steady state is reached. again under the influence of Light radiation

→ Let new values of holes and e⁻ be, Excess holes $= n''_h - n_{ho}$

Excess e⁻ $= n''_e - n_{eo}$

Light radiation creats equal no of electron hole pairs so

$$n''_n - n_{no} = n''_e - n_{eo}$$

So In N-type Ge Bar → e^- ↑ (Large
h ↓ (Negligible

→ So we can say that radiation basically affects the concentration of minority carrier and concentration of majority carrier is hardly affected

→ If light radiation continuously falls on S.C bar, the concentration of e^- or holes will be maintained constant at n_n and n_e respectively. But if radiation are removed, the excess charge carriers will recombine and their concentration will return to original steady state value n_{no} and n_{eo}.

The Light radiation are removed at $t=0$ then excess minority carries delay with time. so

→ Decrease in concentration of holes per second due to recombination

$$= \frac{n_h}{\tau_h} \quad \text{---(1)}$$

At the same time holes are generated due to thermal energy so

→ Increase in concentration of holes per second due to generation $= G$ ---(2)

G → generation rate

→ The rate of increase will be more than rate of decrease before thermal equilibrium reaches. Let the net rate of increase of holes be indicated by $\frac{dn_h}{dt}$, then

$$\frac{dn_h}{dt} = G - \frac{n_h}{\tau_h} \quad \text{---(3)}$$

At $t = t'$ → again steady-state reaches and the concentration of holes becomes constant (equal to n_{h0}) So,

$$\frac{dn_h}{dt} = 0$$

So

$$G = \frac{n_{h0}}{\tau_h}$$

at $t = t'$
$n_h = n_{h0}$

So

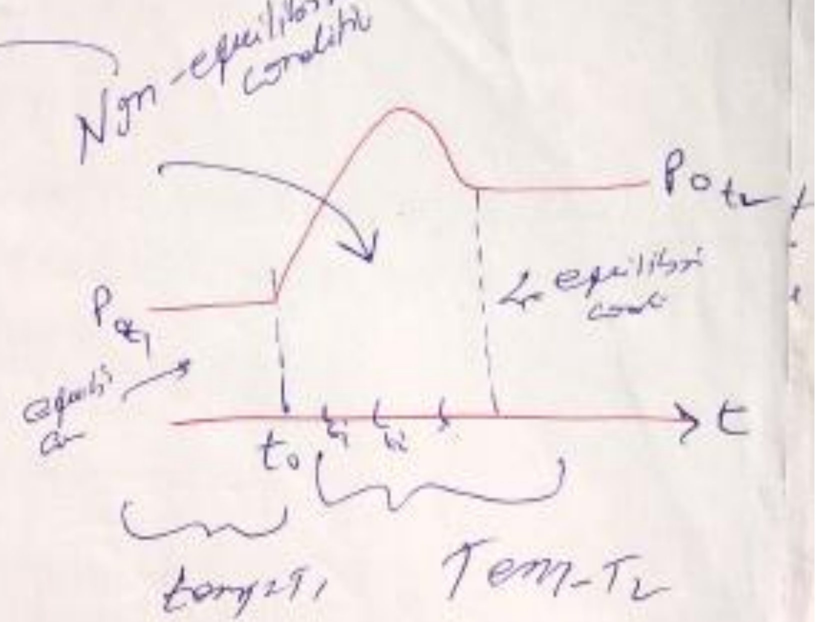
$$\frac{dn_h}{dt} = \frac{n_{h0}}{\tau_h} - \frac{n_h}{\tau_h}$$

$$\frac{dn_h}{dt} = \frac{n_{h0} - n_h}{\tau_h}$$

Where :-

n_h → net (total) holes at any instant

n_{h0} → concentration of holes under steady state with photo excitation



Under this condition
R ≠ G

G > R

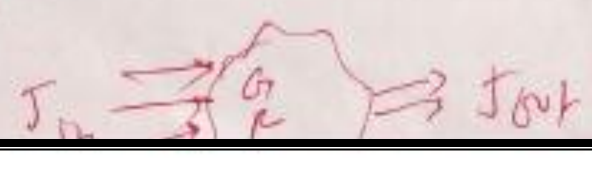
Steady-state — Thermal equilibrium

Mass-Action law is based on thermal equilibrium

So it we have to carrier concentration in non equilibrium

[Net rate of change of carrier conc] = [Generation rate − recombination rate] + [Net carrier flowing in & out of the region of ~~observati~~ observation]

net diff
= G − R



CONTINUITY EQUATION

The **continuity equation** describes a basic concept, namely that a change in carrier density over time is due to the difference between the incoming and outgoing flux of carriers plus the generation and minus the recombination. The flow of carriers and recombination and generation rates.

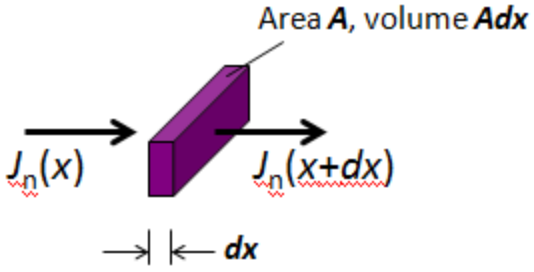
- The ***continuity equations*** are established based on conservation of carriers, and therefore hold generally:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n(x)}{\partial x} - R_n + G_L \qquad \frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_n(x)}{\partial x} - R_p + G_L$$

- The ***minority carrier diffusion equations*** are derived from the continuity equations, specifically for minority carriers under certain conditions (small *E*-field, low-level injection, uniform doping profile):

$$\frac{\partial \Delta n_p}{\partial t} = D_N \frac{\partial^2 \Delta n_p}{\partial x^2} - R_n + G_L \qquad \frac{\partial \Delta p_n}{\partial t} = D_P \frac{\partial^2 \Delta p_n}{\partial x^2} - R_p + G_L$$

- Consider carrier-flux into/out-of an infinitesimal volume:



$$Adx\left(\frac{\partial n}{\partial t}\right) = -\frac{1}{q}\left[J_n(x)A - J_n(x+dx)A\right] - \frac{\Delta n}{\tau_n}Adx$$

$$J_n(x+dx) = J_n(x) + \frac{\partial J_n(x)}{\partial x} dx$$

$$\Rightarrow \quad \frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n(x)}{\partial x} - \frac{\Delta n}{\tau_n}$$

Continuity Equations:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n(x)}{\partial x} - \frac{\Delta n}{\tau_n} + G_L$$

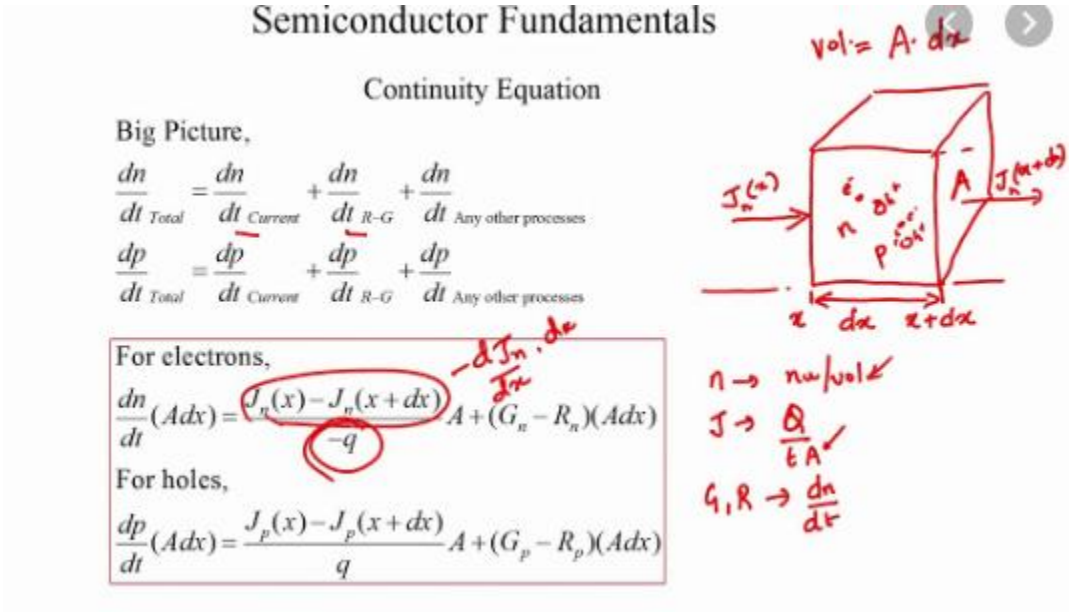
$$\frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_p(x)}{\partial x} - \frac{\Delta p}{\tau_p} + G_L$$

- The ***continuity equations*** are established based on conservation of carriers, and therefore hold generally:

$$\frac{\partial n}{\partial t} = \frac{1}{q}\frac{\partial J_n(x)}{\partial x} - \frac{\Delta n}{\tau_n} + G_L \qquad \frac{\partial p}{\partial t} = -\frac{1}{q}\frac{\partial J_n(x)}{\partial x} - \frac{\Delta p}{\tau_p} + G_L$$

- The ***minority carrier diffusion equations*** are derived from the continuity equations, specifically for minority carriers under certain conditions (small *E*-field, low-level injection, uniform doping profile):

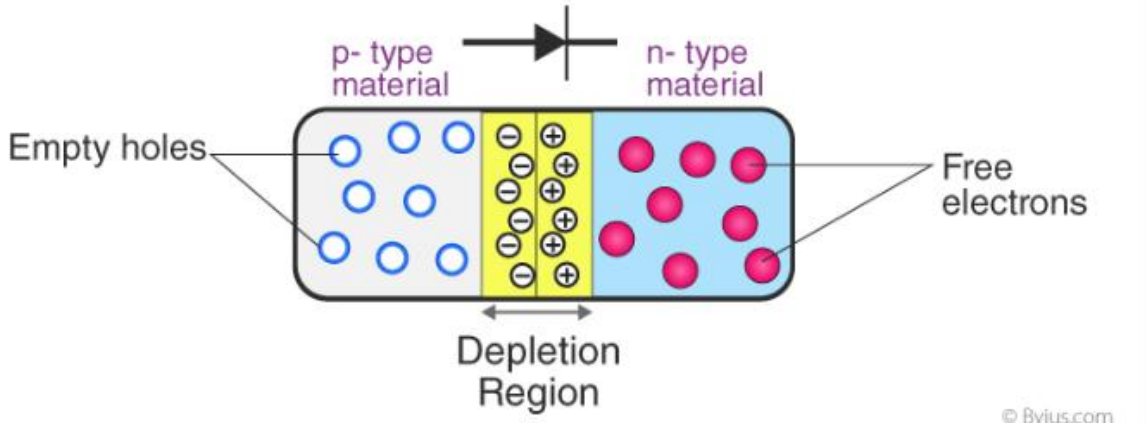
$$\frac{\partial \Delta n_p}{\partial t} = D_N \frac{\partial^2 \Delta n_p}{\partial x^2} - \frac{\Delta n_p}{\tau_n} + G_L \qquad \frac{\partial \Delta p_n}{\partial t} = D_P \frac{\partial^2 \Delta p_n}{\partial x^2} - \frac{\Delta p_n}{\tau_p} + G_L$$



P-N Junction

Definition: ***A p-n junction is an interface or a boundary between two semiconductor material types, namely the p-type and the n-type, inside a semiconductor.***

The p-side or the positive side of the semiconductor has an excess of holes and the n-side or the negative side has an excess of electrons. In a semiconductor, the p-n junction is created by the method of doping. The process of doping is explained in further detail in the next section.



Formation of P-N Junction :

As we know if we use different semiconductor materials to make a p-n junction, there will be a grain boundary that would inhibit the movement of electrons from one side to the other by scattering the electrons and holes and thus we use the process of doping. We will understand the process of doping with the help of this example. Let us consider a thin p-type silicon semiconductor sheet. If we add a small amount of pentavalent impurity to this, a part of the p-type Si will get converted to n-type silicon. This sheet will now contain both p-type region and n-type region and a junction between these two regions. The processes that follow after the formation of a p-n junction are of two types – diffusion and drift. As we know, there is a difference in the concentration of holes and electrons at the two sides of a junction, the holes from the p-side diffuse to the n-side and the electrons from the n-side diffuse to the p-side. These give rise to a diffusion current across the junction.

Also, when an electron diffuses from the n-side to the p-side, an ionized donor is left behind on the n-side, which is immobile. As the process goes on, a layer of positive charge is developed on the n-side of the junction. Similarly, when a hole goes from the p-side to the n-side, and ionized acceptor is left behind in the p-side, resulting in the formation of a layer of negative charges in the p-side of the junction. This region of positive charge and negative charge on either side of the junction is termed as the depletion region. Due to this positive space charge region on either side of the junction, an electric field direction from a positive charge towards the negative charge is developed. Due to this electric field, an electron on the p-side of the junction moves to the n-side of the junction. This motion is termed as the drift. Here, we see that the direction of drift current is opposite to that of the diffusion current.

Biasing conditions for the p-n Junction Diode :

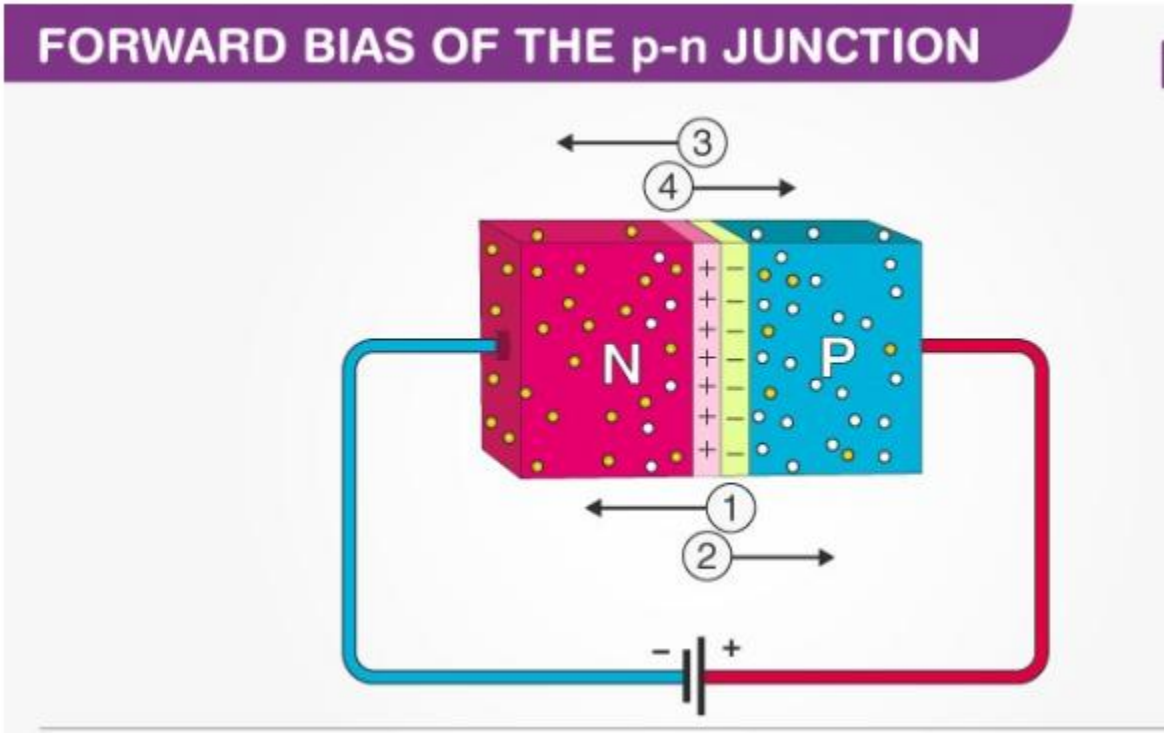
There are two operating regions in the p-n junction diode:

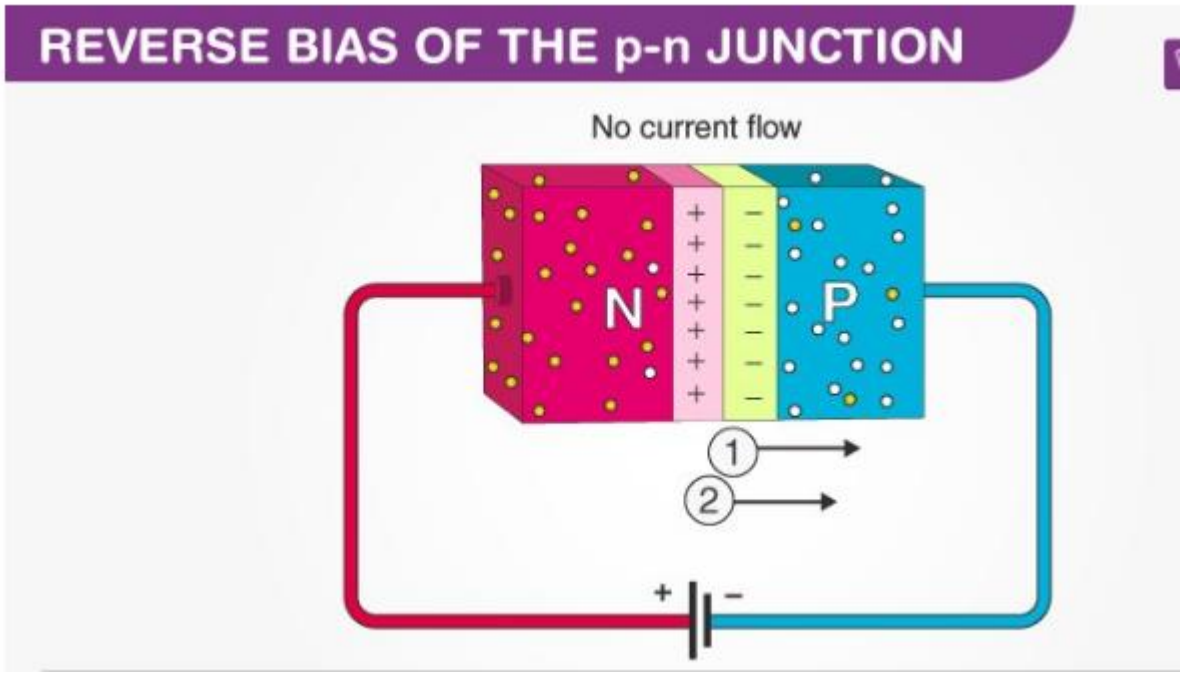
- P-type
- N-type

There are three biasing conditions for p-n junction diode and this is based on the voltage applied:

- Zero bias: There is no external voltage applied to the p-n junction diode.

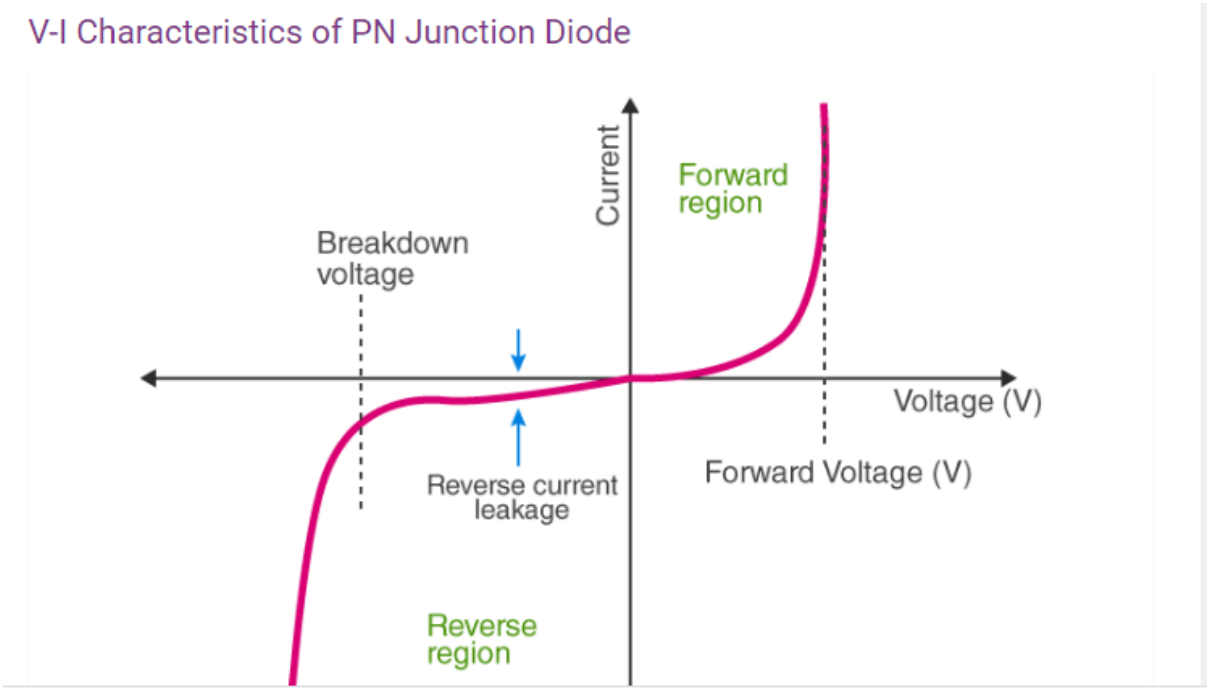
- Forward bias: The positive terminal of the voltage potential is connected to the p-type while the negative terminal is connected to the n-type.
- Reverse bias: The negative terminal of the voltage potential is connected to the p-type and the positive is connected to the n-type.





Current flow in PN junction diode :

The flow of electrons from the n-side towards the p-side of the junction takes place when there is an increase in the voltage. Similarly, the flow of holes from the p-side towards the n-side of the junction takes place along with the increase in the voltage. This results in the concentration gradient between both sides of the terminals. Because of the formation of the concentration gradient, there will be a flow of charge carriers from higher concentration regions to lower concentration regions. The movement of charge carriers inside the pn junction is the reason behind the current flow in the circuit.



VI characteristics of PN junction diode is a curve between the voltage and current through the circuit. Voltage is taken along the x-axis while the current is taken along the y-axis. The above graph is the VI characteristics curve of the PN junction diode. With the help of the curve we can understand that there are three regions in which the diode works, and they are:

- Zero bias
- Forward bias
- Reverse bias

When the PN junction diode is under zero bias condition, there is no external voltage applied and this means that the potential barrier at the junction does not allow the flow of current.

When the PN junction diode is under forward bias condition, the p-type is connected to the positive terminal while the n-type is connected to the negative terminal of the external voltage. When the diode is arranged in this manner, there is a reduction in the potential barrier. For silicone diodes, when the voltage is 0.7 V and for germanium diodes, when the voltage is 0.3 V, the potential barriers decreases and there is a flow of current.

When the diode is in forward bias, the current increases slowly and the curve obtained is non-linear as the voltage applied to the diode is overcoming the potential barrier. Once the potential barrier is overcome by the diode, the diode behaves normal and the curve rises sharply as the external voltage increases and the curve so obtained is linear.

When the PN junction diode is under negative bias condition, the p-type is connected to the negative terminal while the n-type is connected to the positive terminal of the external voltage. This results in an increase in the potential barrier. Reverse saturation current flows in the beginning as minority carriers are present in the junction.

When the applied voltage is increased, the minority charges will have increased kinetic energy which affects the majority charges. This is the stage when the diode breaks down. This may also destroy the diode.

Applications of PN Junction Diode

- p-n junction diode can be used as a photodiode as the diode is sensitive to the light when the configuration of the diode is reverse-biased.
- It can be used as a solar cell.
- When the diode is forward-biased, it can be used in LED lighting applications.
- It is used as rectifiers in many electric circuits and as a voltage-controlled oscillator in varactors.

Zener Diode :

A Zener diode not only allows current to flow from anode to cathode but also, in the reverse direction on reaching the Zener voltage. Due to this functionality, Zener diodes are the most commonly used semiconductor diodes. In this article, let us learn the function of Zener diodes along with its construction, operation and more.

Zener Diode Explanation

A Zener Diode, also known as a breakdown diode, is a heavily doped semiconductor device that is designed to operate in the reverse direction. When the voltage across the terminals of a Zener diode is reversed and the potential reaches the Zener Voltage (knee voltage), the junction breaks down and the current flows in the reverse direction. This effect is known as the Zener Effect.

Zener Diode Definition

A Zener diode is a heavily doped semiconductor device that is designed to operate in the reverse direction. Zener diodes are manufactured with a great variety of Zener voltages (Vz) and some are even made variable.

How does a Zener Diode work in reverse bias?

A Zener diode operates just like a normal diode when it is forward-biased. However, when connected in reverse biased mode, a small leakage current flows through the diode. As the reverse voltage increases to the predetermined breakdown voltage (Vz), current starts flowing through the diode. The current increases to a maximum, which is determined by the series resistor, after which it stabilizes and remains constant over a wide range of applied voltage.

There are two types of breakdowns for a Zener Diode:

- **Avalanche Breakdown**
- **Zener Breakdown**

Avalanche Breakdown in Zener Diode

Avalanche breakdown occurs both in normal diode and Zener Diode at high reverse voltage. When a high value of reverse voltage is applied to the PN junction, the free electrons gain sufficient energy and accelerate at high velocities. These free electrons moving at high velocity collides other atoms and knocks off more electrons. Due to this continuous collision, a large number of free electrons are generated as a result of electric current in the diode rapidly increases. This sudden increase in electric current may permanently destroy the normal diode, however, a Zener diode is designed to operate under avalanche breakdown and can sustain the sudden spike of current. Avalanche breakdown occurs in Zener diodes with Zener voltage (Vz) greater than 6V.

Zener Breakdown in Zener Diode

When the applied reverse bias voltage reaches closer to the Zener voltage, the electric field in the depletion region gets strong enough to pull electrons from their valence band. The valence electrons that gain sufficient energy from the strong electric field of the depletion region break free from the parent atom. At the Zener breakdown region, a small increase in the voltage results in the rapid increase of the electric current.

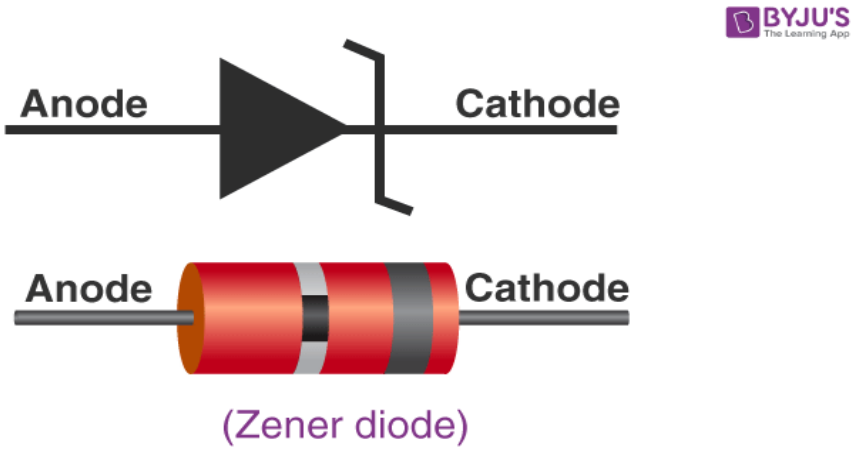
Avalanche Breakdown vs Zener Breakdown

- The Zener effect is dominant in voltages up to 5.6 volts and the avalanche effect takes over above that.
- They are both similar effects, the difference being that the Zener effect is a quantum phenomenon and the avalanche effect is the movement of electrons in the valence band like in any electric current.
- Avalanche effect also allows a larger current through the diode than what a Zener breakdown would allow.

Read More: Difference Between Zener Breakdown and Avalanche Breakdown

Circuit Symbol of Zener Diode

There are many ways in which a Zener diode is packaged. Some are used for high levels of power dissipation and the others are contained with surface mount formats. The most common type of Zener diode is contained within a small glass encapsulation. It has a band around one end marking the cathode side of the diode.



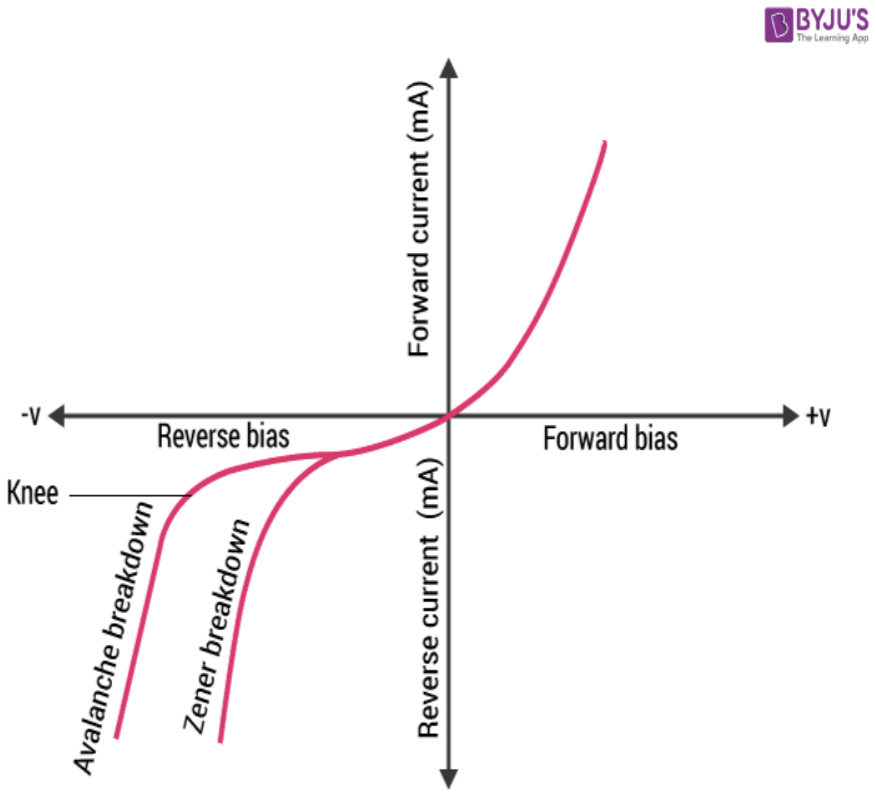
Zener diode symbol and package outlines

From the diagram, we can see that the band around the package corresponds to the line on the diode circuit symbol and this can be an easy way of remembering which end is for which.

The Zener diode circuit symbol places two tags at the end of the bar – one in the upward direction and the other in the lower direction as shown in the figure. This helps in distinguishing Zener diodes from other forms of diodes within the circuit.

V-I Characteristics of Zener Diode

The diagram given below shows the V-I characteristics of the Zener diode.



When reverse-biased voltage is applied to a Zener diode, it allows only a small amount of leakage current until the voltage is less than Zener voltage.

The V-I characteristics of a Zener diode can be divided into two parts as follows:

(i) Forward Characteristics

(ii) Reverse Characteristics

Forward Characteristics of Zener Diode

The first quadrant in the graph represents the forward characteristics of a Zener diode. From the graph, we understand that it is almost identical to the forward characteristics of any other P-N junction diode.

Reverse Characteristics of Zener Diode

When a reverse voltage is applied to a Zener voltage, initially a small reverse saturation current Io flows across the diode. This current is due to thermally generated minority carriers. As the reverse voltage is increased, at a certain value of reverse voltage, the reverse current increases drastically and sharply. This is an indication that the breakdown has occurred. We call this voltage breakdown voltage or Zener voltage and it is denoted by Vz.

Zener Diode Specifications

Some commonly used specifications for Zener diodes are as follows:

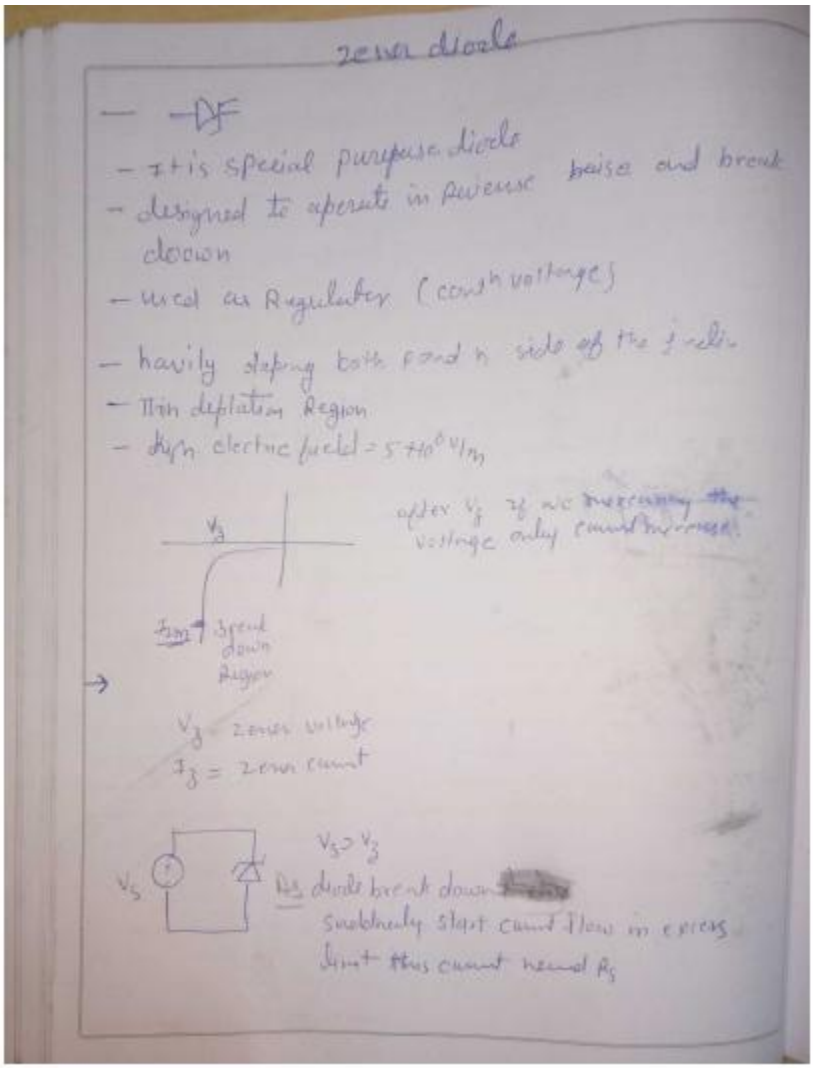
- **Zener/Breakdown Voltage** – The Zener or the reverse breakdown voltage ranges from 2.4 V to 200 V, sometimes it can go up to 1 kV while the maximum for the surface-mounted device is 47 V.
- **Current Iz (max)** – It is the maximum current at the rated Zener Voltage (Vz – 200μA to 200 A)
- **Current Iz (min)** – It is the minimum value of current required for the diode to breakdown.
- **Power Rating** – It denotes the maximum power the Zener diode can dissipate. It is given by the product of the voltage of the diode and the current flowing through it.
- **Temperature Stability** – Diodes around 5 V have the best stability
- **Voltage Tolerance** – It is typically ±5%
- **Zener Resistance (Rz)** – It is the resistance to the Zener diode exhibits.

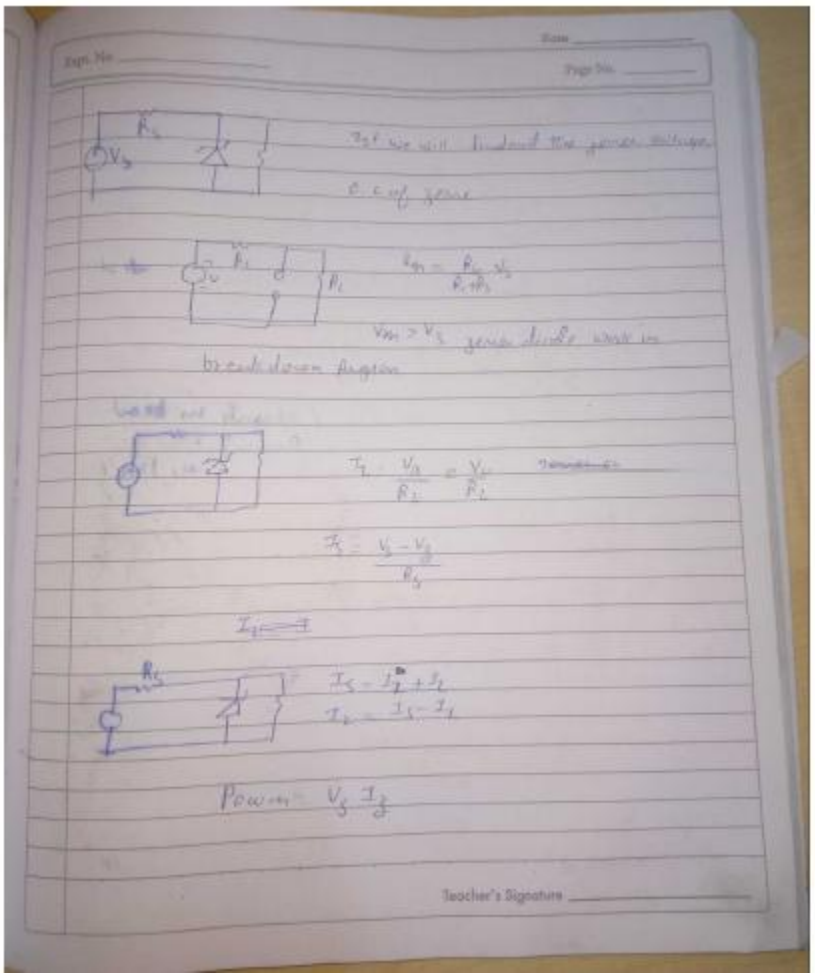
Application of Zener Diode

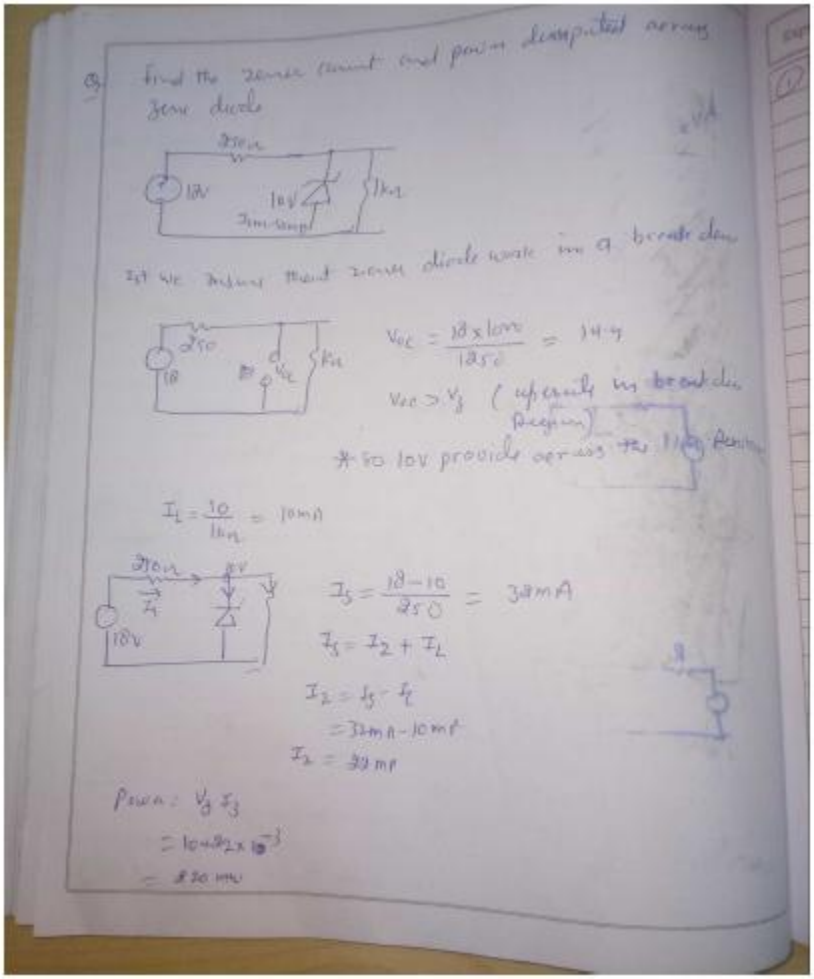
Following are the applications of Zener diode:

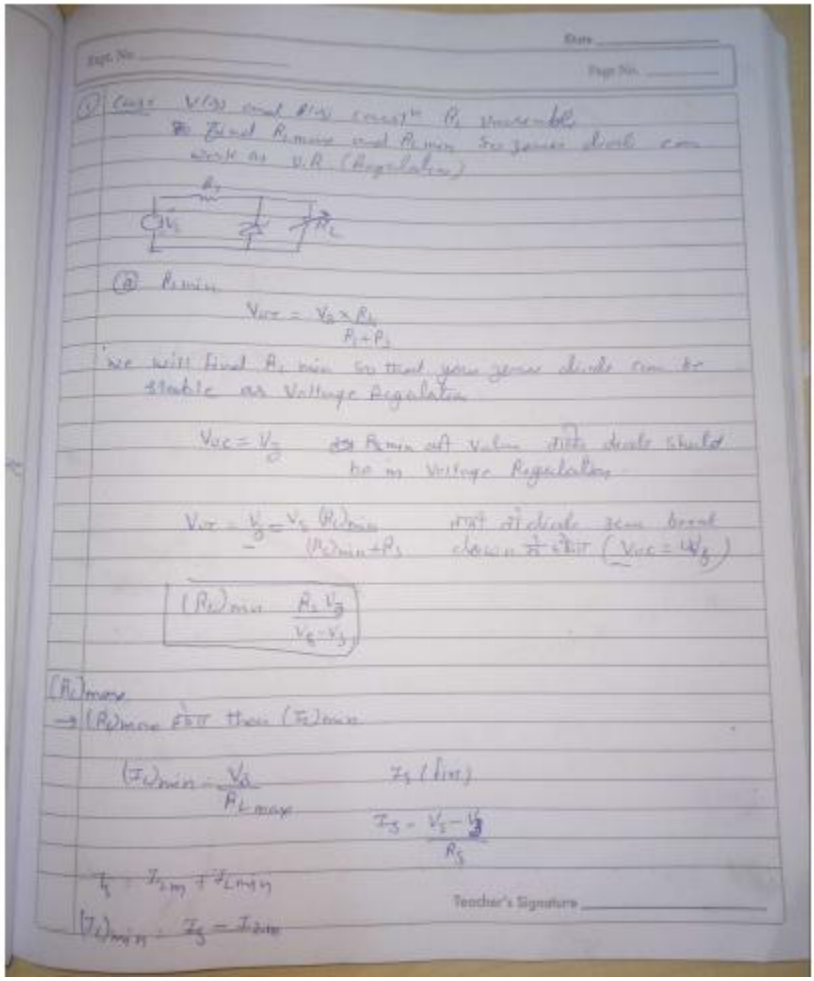
Zener diode as a voltage regulator:

Zener diode is used as a Shunt voltage regulator for regulating voltage across small loads. The breakdown voltage of Zener diodes will be constant for a wide range of current. Zener diode is connected parallel to the load to make it reverse bias and once the Zener diode exceeds knee voltage, the voltage across the load will become constant.

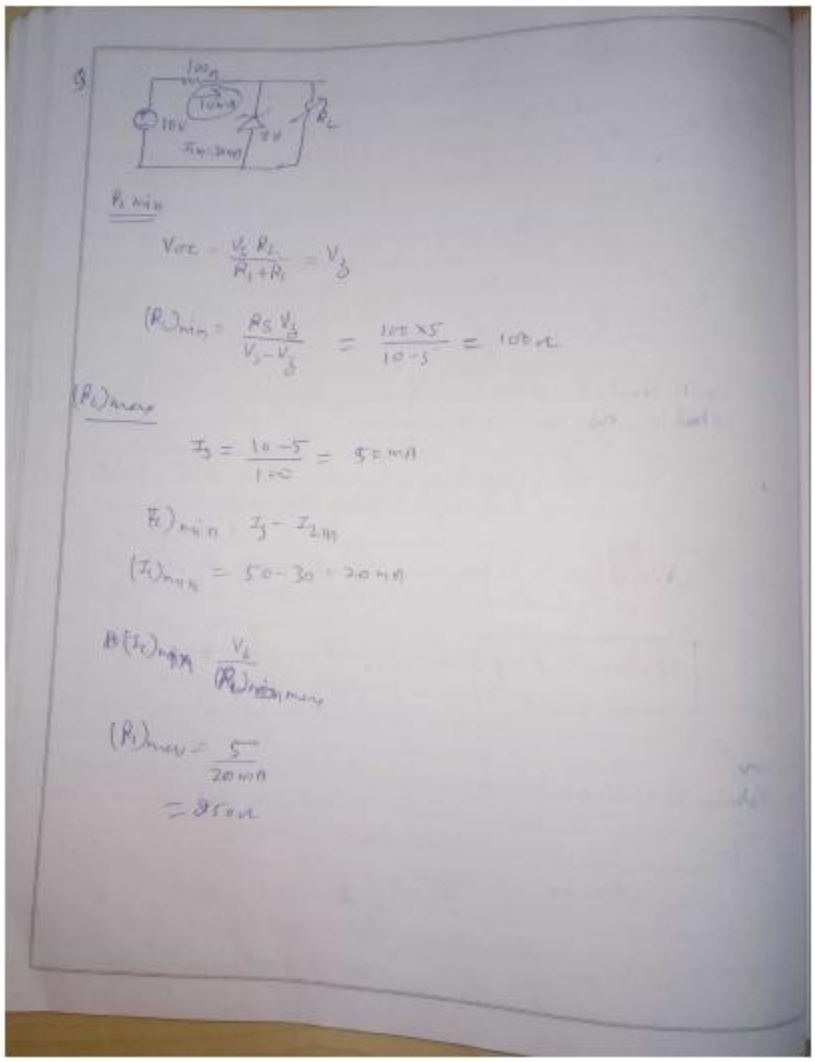

zener diode
- It is special purpose diode
- designed to operate in reverse biase and break down
- used as Regulator (constant voltage)
- heavily doping both p and n side of the junction
- Thin depletion Region
- High electric field = 5×10^5 V/m
after V_z if we increase the voltage only current increase
Zener
Break down Region
V_z = zener voltage
I_z = zener current
V_s
$V_s > V_z$
Zener diode break down
Suddenly start current flow in excess
limit this current need R_s

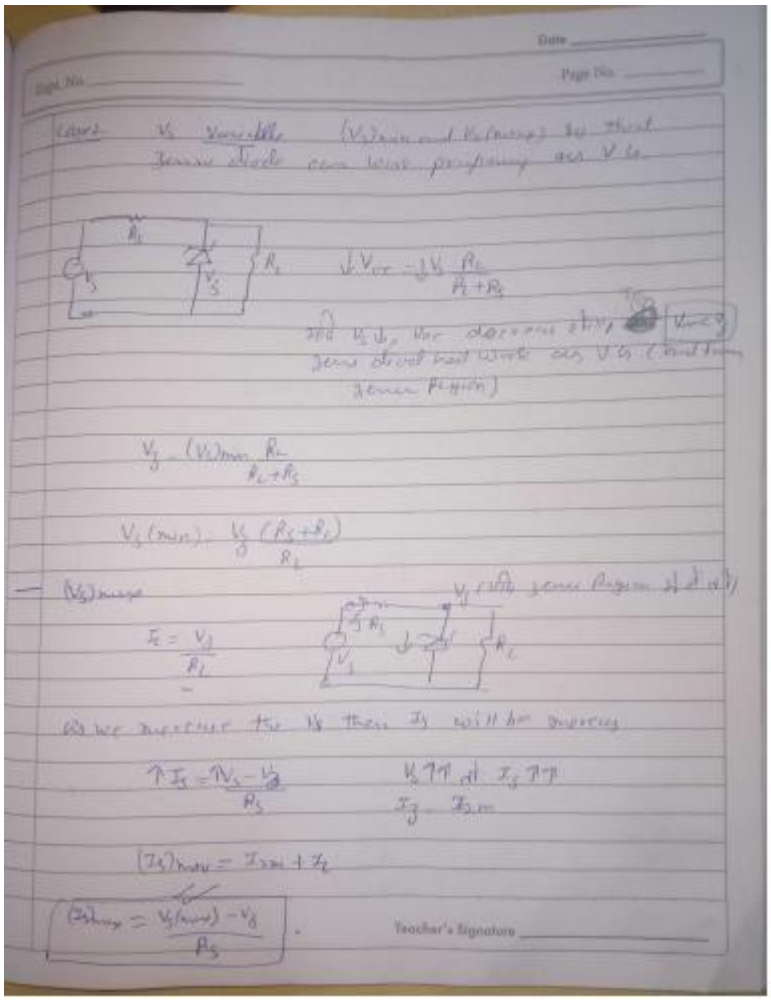

Teacher's Signature


Find the zener current and power dissipated across zener diode
250Ω
18V
10V
1kΩ
7mA-amp
If we assume that zener diode work in a breakdown
250
18
1kΩ
$V_{cc} = \frac{18 \times 1000}{1250} = 14.4$
$V_{cc} > V_z$ (zener in breakdown region)
So 10V provide across
$I_L = \frac{10}{1k\Omega} = 10mA$
250Ω
18V
$I_S = \frac{18-10}{250} = 32mA$
$I_S = I_Z + I_L$
$I_Z = I_S - I_L$
$= 32mA - 10mA$
$I_Z = 22mA$
Power: $V_Z I_Z$
$= 10 \times 22 \times 10^{-3}$
$= 220mW$



Teacher's Signature _____





Case 2 V_s Variable (V_s)min and (V_s)max to find Zener diode can work properly acc. V_s to

$V_{oc} = \frac{V_s R_L}{R_1 + R_L}$

If $V_s \downarrow$, then decrease etc, $V_o = 0$ Zener diode not work as V_s (condition Zener Region)

$V_z = (V_s)_{min} \frac{R_L}{R_L + R_s}$

$V_s(min) = \frac{V_z (R_s + R_L)}{R_L}$

$(V_s)_{max}$ V_z in Zener Region

$I_z = \frac{V_z}{R_L}$

As we increase the V_s then I_s will be increase

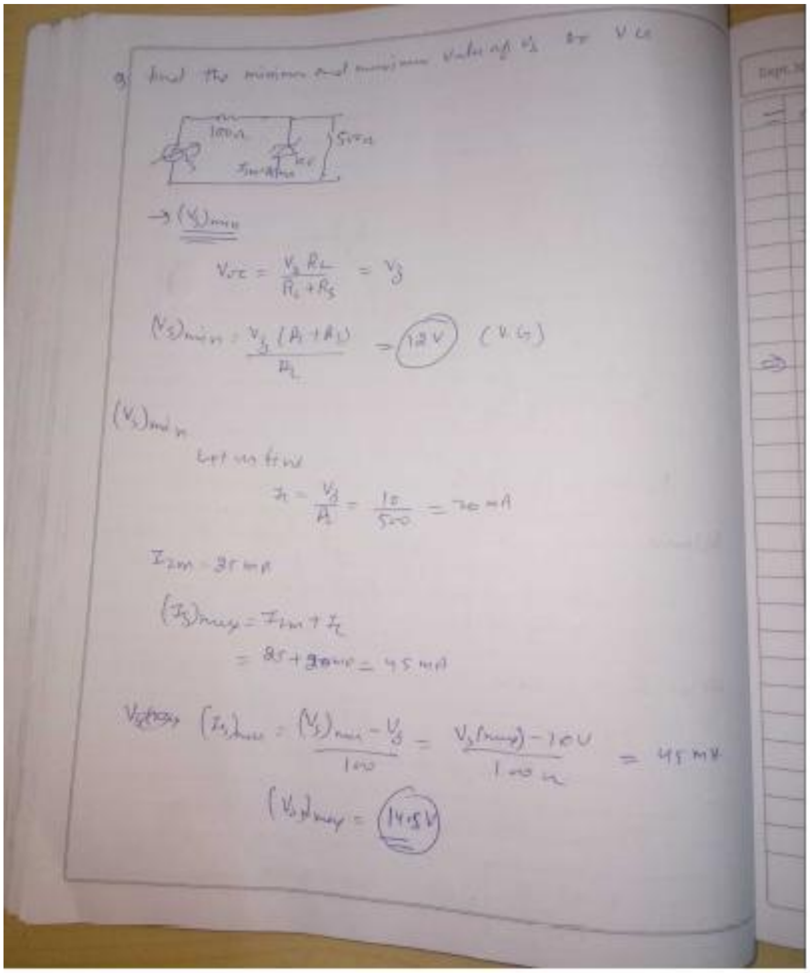
$I_s = \frac{V_s - V_z}{R_s}$ $V_s \uparrow$ & $I_s \uparrow$

$I_z = I_{zm}$

$(I_s)_{max} = I_{zm} + I_L$

$(I_s)_{max} = \frac{V_s(max) - V_z}{R_s}$

Teacher's Signature



Q) Find the minimum and maximum value of V_s for V_Z

→ $(V_S)_{min}$

$V_{Z} = \frac{V_s R_L}{R_L + R_S} = V_Z$

$(V_S)_{min} = \frac{V_Z (R_L + R_S)}{R_L} = 12V$ (V.G)

$(V_S)_{max}$

Let us find

$I_L = \frac{V_Z}{R_L} = \frac{10}{500} = 20mA$

$I_{ZM} = 25mA$

$(I_S)_{max} = I_{ZM} + I_L$

$= 25 + 20mA = 45mA$

$(I_S)_{max} = \frac{(V_S)_{max} - V_Z}{100} = \frac{V_S(max) - 10V}{100\Omega} = 45mA$

$(V_S)_{max} = 14.5V$

Schottky Diode

What is Schottky Diode?

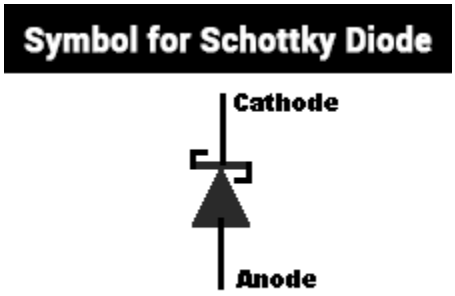
The **Schottky Diode** is an electronic component that is used to radio frequency (RF) functions like a mixer or a detector diode. It comes useful for functions having power because of possessing low forward voltage drop that would result in power loss at the least levels in comparison to the general PN junction diodes. In some cases, it is sometimes addressed as the surface barrier diode or hot carrier diode.

The other names of a Schottky diode are barrier diode, hot-electron diode, hot carrier diode, majority carrier diode.

SymbolV-I CharacteristicsWorkingApplicationsAdvantagesDisadvantagesDifference

What is Schottky Diode Symbol?

The symbol for the Schottky barrier diode is based around the basic diode circuit symbol. The circuit symbol of the Schottky diode is shown in the figure.

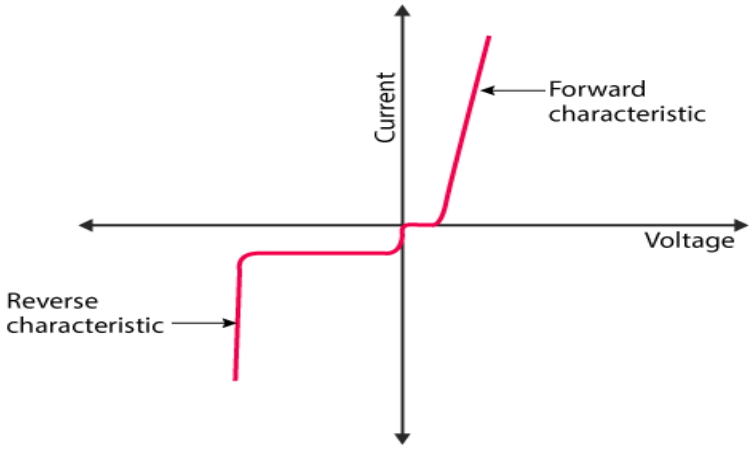


Schottky Diode Symbol

V-I Characteristics of Schottky Diode

The V-I characteristics of Schottky diode are very much similar to the PN junction diode. Current is the dependent variable while voltage is the independent variable in the Schottky diode. The forward voltage drop of

the Schottky diode is low between 0.2 to 0.3 volts.

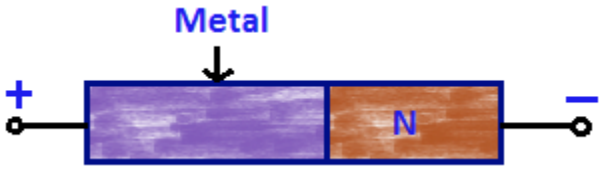


You may also want to check out these topics given below!

- Semiconductor Diode
- Laser Diode
- Uses of Diode

Working of a Schottky Diode

- The operation relies on the principle that the electrons in different materials have different potential energy.
- N-type semiconductors have higher potential energy than electrons of metals.



- When these two are brought into contact, there is a flow of electrons in both directions across the metal-semiconductor interface.
- A voltage is applied to the Schottky so that the metal is positive when compared to the semiconductor.
- The voltage opposes the built-in potential and makes the current flow easy.

Applications of Schottky Diode

Schottky diodes have been useful for the industry of electronics that has spotted many applications in diode rectifiers because of its unique properties. Here are some major areas where it is widely used.

RF mixer and detector diode:

The Schottky diode consists of its radio frequency functions owing to its switching speed at the highest level and top frequency capability. The Schottky barrier diodes come handy for diode ring mixers with high performance.

Power rectifier:

The Schottky barrier diodes also have functions with high power as rectifiers. The high density of current and voltage drop with low forward shows that the wastage of power is lesser than the normal PN junction diodes.

Power OR circuits:

This diode would be useful for functions where two different power supplies drive a load like in battery supply. It is important that the power coming from supply should not mix with the others.

Solar Cell Applications:

As we know, the solar cells are usually linked to the batteries that are rechargeable, mostly batteries with lead-acid since power supply must be necessary round the clock. Solar cells would not support the applied charge in reverse and thus, a diode would be used in a proportional pattern of the solar cells.

Advantages of Schottky diode

Following are the advantages of Schottky diode:

- The capacitance of the diode is low as the depletion region of the diode is negligible.
- The reverse recovery time of the diode is very fast, that is the change from ON to OFF state is fast.
- The current density of the diode is high as the depletion region is negligible.
- The turn-on voltage of the diode is 0.2 to 0.3 volts, which is very low.

Disadvantages of Schottky diode

The only disadvantage of Schottky diode is that the reverse saturation current of the diode is large.

What is the difference between Schottky diode and PN junction diode?

Schottky diode	**PN junction diode**
In this diode, the junction is formed between the n-type semiconductor and	In this diode, the junction is formed between the p-type and n-type

the metal plate	semiconductors
The forward voltage drop is low	The forward voltage drop for pn junction diode is more
Reverse recovery loss and reverse recovery time are very less	Reverse recovery loss and reverse recovery time are more
It is a unipolar device	It is a bipolar device
The conduction of current happens only due to the movement of electrons	The conduction of current happens due to the movement of electrons and holes

Unit 4

Biploar transistor:

A transistor is basically a Si on Ge crystal containing three separate regions. It can be either NPN or PNP type. The middle region is called the base and the outer two regions are called emitter and the collector. The outer layers although they are of same type but their functions cannot be changed. They have different physical and electrical properties.

In most transistors, emitter is heavily doped. Its job is to emit or inject electrons into the base. These bases are lightly doped and very thin, it passes most of the emitter-injected electrons on to the collector. The doping level of collector is intermediate between the heavy doping of emitter and the light doping of the base.

The collector is so named because it collects electrons from base. The collector is the largest of the three regions; it must dissipate more heat than the emitter or base. The transistor has two junctions. One between emitter and the base and other between the base and the collector. Because of this the transistor is similar to two diodes, one emitter diode and other collector base diode.

When transistor is made, the diffusion of free electrons across the junction produces two depletion layers. For each of these depletion layers, the barrier potential is 0.7 V for Si transistor and 0.3 V for Ge transistor.

The depletion layers do not have the same width, because different regions have different doping levels. The more heavily doped a region is, the greater the concentration of ions near the junction. This means the depletion layer penetrates more deeply into the base and slightly into emitter. Similarly, it penetration more into collector. The thickness of collector depletion layer is large while the base depletion layer is small as shown in

both the junctions are forward biased using two d.c sources, as shown in free electrons (majority carriers) enter the emitter and collector of the transistor, joins at the base and come out of the base. Because both the diodes are forward biased, the emitter and collector currents are large**.** If both the junction are reverse biased as shown in then small currents flows through both junctions only due to thermally produced minority carriers

and surface leakage. Thermally produced carriers are temperature dependent it approximately doubles for every 10 degree celsius rise in ambient temperature. The surface leakage current increases with voltage**.**

When the emitter diode is forward biased and collector diode is reverse biased as shown in **fig.** 3.1 then one expect large emitter current and small collector current but collector current is almost as large as emitter current.

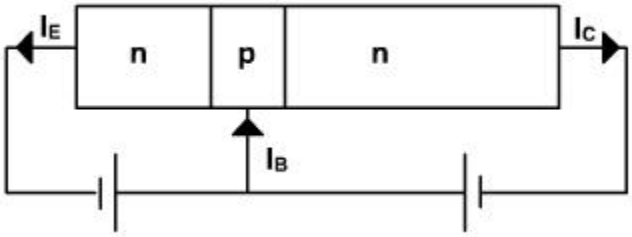


Fig.3.1

When emitter diodes forward biased and the applied voltage is more than 0.7 V (barrier potential) then larger number of majority carriers (electrons in n-type) diffuse across the junction.

Once the electrons are injected by the emitter enter into the base, they become minority carriers. These electrons do not have separate identities from those, which are thermally generated, in the base region itself. The base is made very thin and is very lightly doped. Because of this only few electrons traveling from the emitter to base region recombine with holes. This gives rise to recombination current. The rest of the electrons exist for more time. Since the collector diode is reverse biased, (n is connected to positive supply) therefore most of the electrons are pushed into collector layer. These collector elections can then flow into the external collector lead.

Thus, there is a steady stream of electrons leaving the negative source terminal and entering the emitter region. The V_{EB} forward bias forces these emitter electrons to enter the base region. The thin and lightly doped base gives almost all those electrons enough lifetime to diffuse into the depletion layer. The depletion

layer field pushes a steady stream of electron into the collector region. These electrons leave the collector and flow into the positive terminal of the voltage source. In most transistor, more than 95% of the emitter injected electrons flow to the collector, less than 5% fall into base holes and flow out the external base lead. But the collector current is less than emitter current.

3.2 Relation between different currents in a transistor:

The total current flowing into the transistor must be equal to the total current flowing out of it. Hence, the emitter current I_E is equal to the sum of the collector (I_C) and base current (I_B). That is,

$$I_E = I_C + I_B \qquad \text{(E-1)}$$

The currents directions are positive directions. The total collector current I_C is made up of two components.

1. The fraction of emitter (electron) current which reaches the collector ($\alpha_{dc} I_E$)

2. The normal reverse leakage current I_{CO}

$$\therefore I_C = \alpha_{dc} I_E + I_{co}$$

$$\text{or} \quad \alpha_{dc} = \frac{I_C - I_{co}}{I_E} \qquad \text{(E-2)}$$

α_{dc} is known as large signal current gain or dc alpha. It is always positive. Since collector current is almost equal to the I_E therefore $\alpha dc\ I_E$ varies from 0.9 to 0.98. Usually, the reverse leakage current is very small compared to the total collector current.

$$\text{Neglecting } I_{co}, \quad \alpha_{dc} = \frac{I_C}{I_E} \qquad \text{(E-3)}$$

NOTE:

The forward bias on the emitter diode controls the number of free electrons infected into the base. The larger (V_{BE}) forward voltage, the greater the number of injected electrons. The reverse bias on the collector diode has little influence on the number of electrons that enter the collector. Increasing V_{CB} does not change the number of free electrons arriving at the collector junction layer.

The symbol of npn and pnp transistors are shown in **fig.** 3.2.

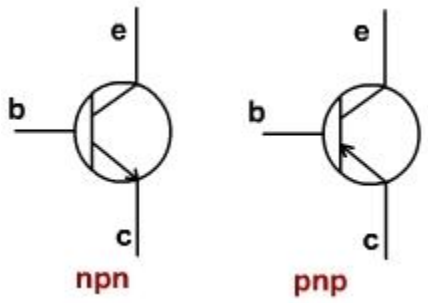


Fig. 3.2

3.3 Breakdown Voltages:

Since the two halves of a transistor are diodes, two much reverse voltage on either diode can cause breakdown. The breakdown voltage depends on the width of the depletion layer and the doping levels. Because of the heavy doping level, the emitter diode has a low breakdown voltage approximately 5 to 30 V. The collector diode is less heavily doped so its breakdown voltage is higher around 20 to 300 V.

3.4 The Common Base Configuration :

If the base is common to the input and output circuits, it is known as common base configuration as shown in **fig.** 3.3.

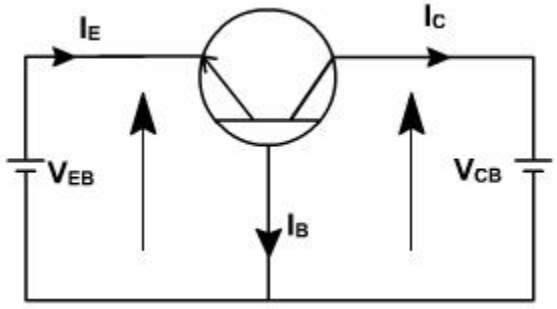


Fig. 3.3

For a pnp transistor the largest current components are due to holes. Holes flow from emitter to collector and few holes flow down towards ground out of the base terminal. The current directions are shown in **fig.** 3.3.

($I_E = I_C + I_B$). **(E-4)**

For a forward biased junction, V_{EB} is positive and for a reverse biased junction V_{CB} is negative. The complete transistor can be described by the following two relations, which give the input voltage V_{EB} and output current I_C in terms of the output voltage (V_{CB}) and input current I_E.

$V_{EB} = f_1(V_{CB}, I_E)$ **(E-5)**

$I_C = f_2(V_{CB}, I_E)$ **(E-6)**

3.4.1The output characteristic:

The collector current I_C is completely determined by the input current I_E and the V_{CB} voltage. The relationship is given in **fig.**3.4**.** It is a plot of I_C versus V_{CB}, with emitter current I_E as parameter. The curves are known as the output or collector or static characteristics. The transistor consists of two diodes placed in series back to back (with two cathodes connected together). The complete characteristic can be divided in three regions.

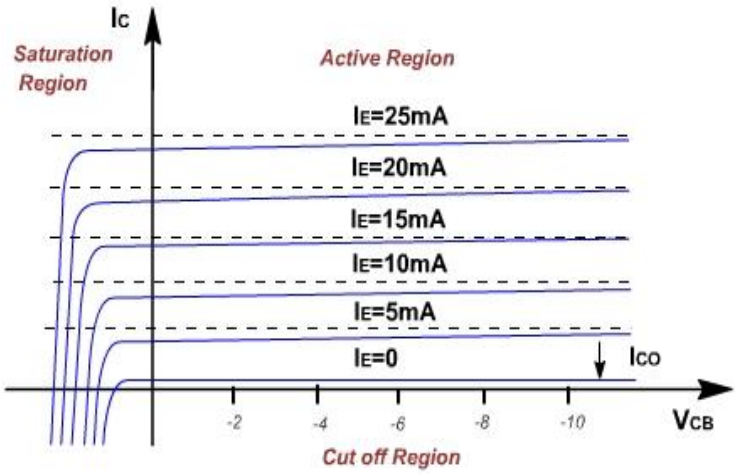


Figure 3.4

(1). Active region:

In this region the collector diode is reverse biased and the emitter diode is forward biased. Consider first that the emitter current is zero. Then the collector current is small and equals the reverse saturation current I_{CO} of the collector junction considered as a diode.

If the forward current I_B is increased, then a fraction of I_E ie. $\alpha_{dc}I_E$ will reach the collector. In the active region, the collector current is essentially independent of collector voltage and depends only upon the emitter current. Because α_{dc} is, less than one but almost equal to unity, the magnitude of the collector current is slightly less that of emitter current. The collector current is almost constant and work as a current source.

The collector current slightly increases with voltage. This is due to early effect. At higher voltage collector gathers in a few more electrons. This reduces the base current. The difference is so small, that it is usually neglected. If the collector voltage is increased, then space charge width increases; this decreased the effective base width. Then there is less chance for recombination within the base region.

(2). Saturation region:

The region to the left of the ordinate V_{CB} = 0, and above the I_E = 0, characteristic in which both emitter and collector junction are forward biased, is called saturation region.

When collector diode is forward biased, there is large change in collector current with small changes in collector voltage. A forward bias means, that p is made positive with respect to n, there is a flow of holes from p to n. This changes the collector current direction. If diode is sufficiently forward biased the current changes rapidly. It does not depend upon emitter current.

(3). Cut off region:

The region below I_E = 0 and to the right of V_{CB} for which emitter and collector junctions are both reversed biased is referred to cutoff region. The characteristics I_E = 0, is similar to other characteristics but not coincident with horizontal axis. The collector current is same as I_{CO}. I_{CBO} is frequently used for I_{CO}. It means collector to base current with emitter open. This is also temperature dependent.

3.4.2 The Input Characteristic:

In the active region the input diode is forward biased, therefore, input characteristic is simply the forward biased characteristic of the emitter to base diode for various collector voltages. **fig. 3.5.** Below cut in voltage (0.7 or 0.3) the emitter current is very small. The curve with the collector open represents the forward biased emitter diode. Because of the early effect the emitter current increases for same V_{EB}. (The diode becomes better diode).

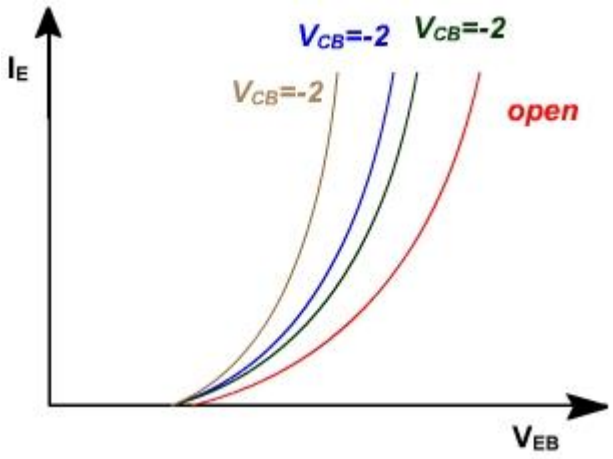


Fig. 3.5

When the collector is shorted to the base, the emitter current increases for a given V_{EB} since the collector now removes minority carriers from the base, and hence base can attract more holes from the emitter. This mean that the curve V_{CB}= 0, is shifted from the character when V_{CB} =

open.

3.5 Common Base Amplifier:

The common base amplifier circuit is shown in **Fig.**3.6**.** The V_{EE} source forward biases the emitter diode and V_{CC} source reverse biased collector diode. The ac source v_{in} is connected to emitter through a coupling capacitor so that it blocks dc. This ac voltage produces small fluctuation in currents and voltages. The load resistance R_L is also connected to collector through coupling capacitor so the fluctuation in collector base voltage will be observed across R_L.

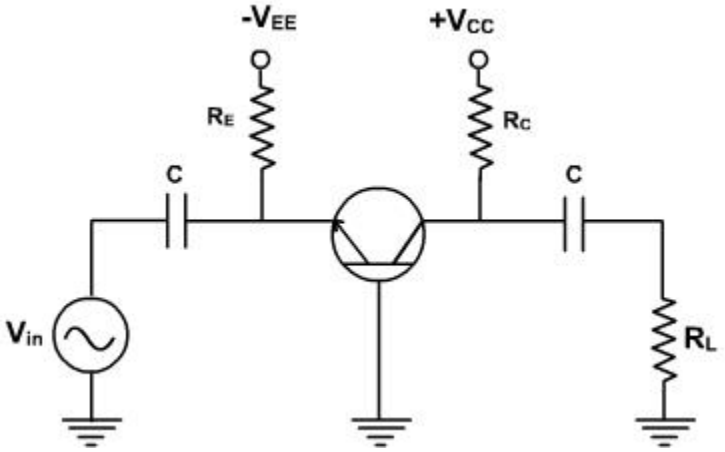


Fig. 3.6

The dc equivalent circuit is obtained by reducing all ac sources to zero and opening all capacitors. The dc collector current is same as I_E and V_{CB} is given by

$V_{CB} = V_{CC} - I_C R_C$. (E-7)

These current and voltage fix the Q point. The ac equivalent circuit is obtained by reducing all dc

sources to zero and shorting all coupling capacitors. r'_e represents the ac resistance of the diode as shown in **Fig.** 3.7.

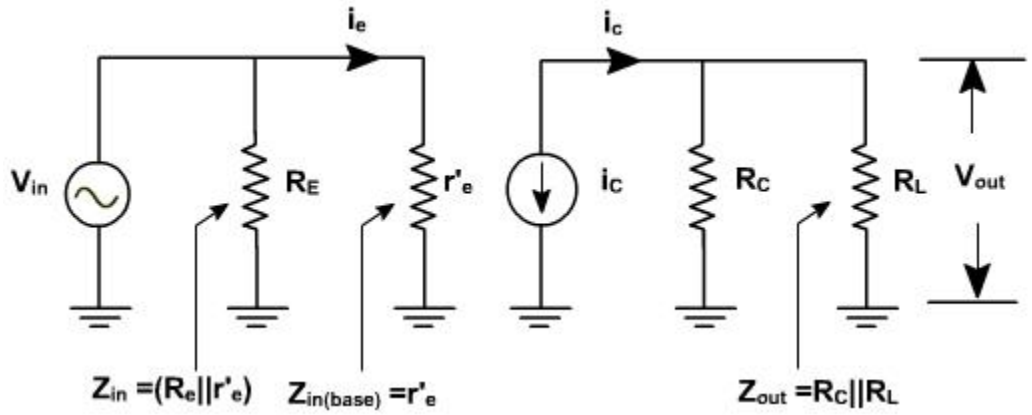


Fig. 3.7

Fig.3.8, shows the diode curve relating I_E and V_{BE}. In the absence of ac signal, the transistor operates at Q point (point of intersection of load line and input characteristic). When the ac signal is applied, the emitter current and voltage also change. If the signal is small, the operating point swings sinusoidally about Q point (A to B).

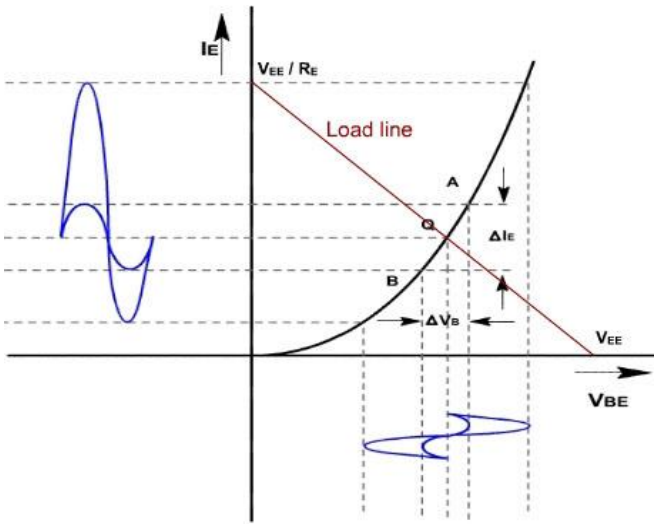


Fig .3.8

If the ac signal is small, the points A and B are close to Q, and arc A B can be approximated by a straight line and diode appears to be a resistance given by

$$r'_e = \left.\frac{\Delta V_{BE}}{\Delta I_E}\right|_{small\ change}$$

$$= \frac{V_{be}}{i_e} = \frac{ac\ voltage\ across\ base\ and\ emitter}{ac\ current\ through\ emitter} \qquad \text{(E-8)}$$

If the input signal is small, input voltage and current will be sinusoidal but if the input voltage is large then current will no longer be sinusoidal because of the non linearity of diode curve. The emitter current is elongated on the positive half cycle and compressed on negative half cycle. Therefore the output will also be distorted.

r'_e is the ratio of ΔV_{BE} and ΔI_E and its value depends upon the location of Q. Higher up the Q point small will be the value of r' e because the same change in V_{BE} produces large change in I_E. The slope of the curve at Q determines the value of r'_e. From calculation it can be proved that.

$r'_e = 25mV / I_E$

Proof

In general, the current through a diode is given by

$$I = I_{CO}\left(e^{\frac{qV}{KT}} - 1\right) \qquad \text{(E-9)}$$

Where q is he charge on electron, V is the drop across diode, T is the temperature and K is a constant.

On differentiating w.r.t V, we get,

$$\frac{dI}{dV} = I_{\infty} * e^{\frac{qV}{KT}} * \frac{q}{KT} \qquad \text{(E-10)}$$

The value of (q / KT) at 25°C is approximately 40.

$$\frac{dI}{dV} = 40 * I_{co} * e^{\frac{qV}{KT}}$$

Therefore, $= 40 * (I + I_{co})$

or, $$\frac{dV}{dI} = \frac{1}{40 * (I + I_{co})} \approx \frac{1}{40 * I} \qquad \text{(E-11)}$$

$$\text{Therefore, ac resistance of the emitter diode} = \frac{dV}{dI} = \frac{25mV}{I} \text{ Ohms}$$

To a close approximation the small changes in collector current equal the small changes in emitter current. In the ac equivalent circuit, the current 'i_C' is shown upward because if 'i_e' increases, then 'i_C' also increases in the same direction.

Voltage gain:

Since the ac input voltage source is connected across r'_e. Therefore, the ac emitter current is given by

$i_e = V_{in} / r'_e$

or, $V_{in} = ie\ r'_e$ (E-12)

The output voltage is given by $V_{out} = i_c (R_C \parallel R_L)$ (E-13)

$$\text{Therefore, voltage gain } A_V = \frac{v_{out}}{v_{in}} = \frac{(R_C \parallel R_L)}{r'_e}$$

$$= \frac{rc}{r} \qquad \text{(E-14)}$$

Under open circuit condition $v_{out} = i_c R_c$ (E-15)

$$\text{Therefore, voltage gain in open circuit condition} = A_V = \frac{R_C}{r'_e} \qquad \text{(E-16)}$$

Example-1

Find the voltage gain and output of the amplifier shown in **fig.3.9**, if input voltage is 1.5mV.

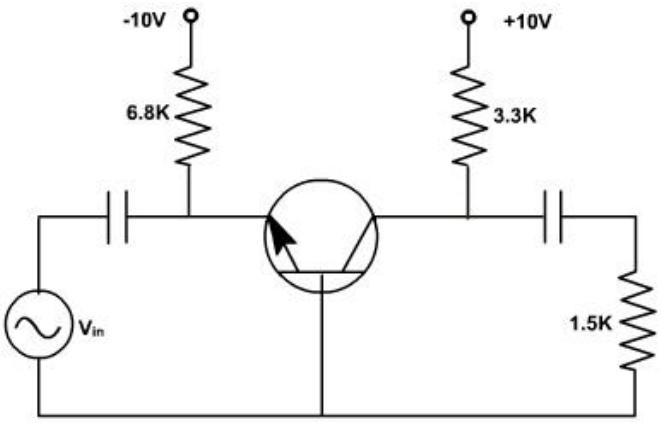


Fig. 3.9

Solution:

The emitter dc current I E is given by $I_E = \frac{10 - 0.7}{6.8k} = 1.37mA$

Therefore, emitter ac resistance = $A_V = \frac{r_c}{r'_e} = \frac{3.3k \| 1.5k}{18.2\Omega}$

or, $A_V = 56.6$

and, $V_{out} = 1.5 \times 56.6 = 84.9$ mV

Example-2

Repeat example-1 if ac source has resistance R s = 100 W .

Solution:

The ac equivalent circuit with ac source resistance is shown in **fig.**3.10.

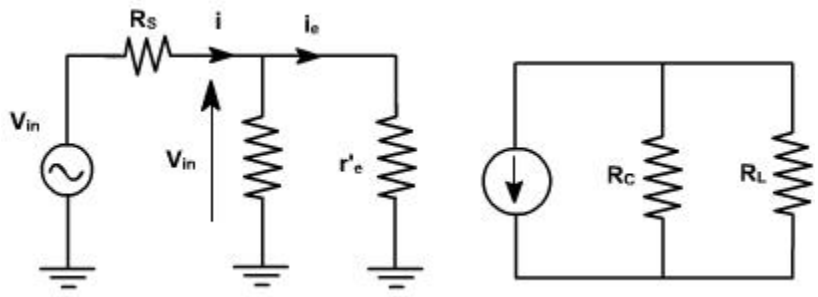


Fig. 3.10

The emitter ac current is given by $i_e = \frac{v_{in}}{R_s + (R_E \| r_e')} \times \frac{R_E}{R_E + r_e'}$

or, $i_e = \frac{v_{in}}{(R_s + r_e')R_E + R_s r_e'} \times R_E \; ; \frac{v_{in}}{R_s + r_e'}$

Therefore, voltage gain of the amplifier = $A_V = \frac{v_{out}}{v_{in}} = \frac{i_c \, r_c}{i_e \, (R_s + r'_e)} = \frac{r_c}{R_s + r_e'}$

$$A_V = \frac{3.3k \| 1.5k}{100\Omega + 18.2\,\Omega} = 8.71$$

and, V_{out} = 1.5 x 8.71 =13.1 mV

3.6 Common Emitter Configuration

Common Emitter Curves:

The common emitter configuration of BJT is shown in **fig. 3.11**.

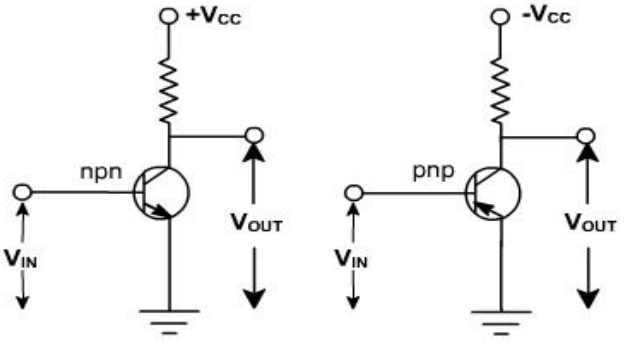


Fig. 3.11

In C.E. configuration the emitter is made common to the input and output. It is also referred to as grounded emitter configuration. It is most commonly used configuration. In this, base current and output voltages are taken as impendent parameters and input voltage and output current as dependent parameters

$V_{BE} = f_1 (I_B, V_{CE})$ (E-17)

$I_C = f_2(I_B, V_{CE})$ (E-18)

3.6.1 Input Characteristic:

The curve between I_B and V_{BE} for different values of V_{CE} are shown in **fig.3.12**. Since the base emitter junction of a transistor is a diode, therefore the characteristic is similar to diode one. With higher values of V_{CE} collector gathers slightly more electrons and therefore base current reduces. Normally this effect is neglected. (Early effect). When collector is shorted with emitter then the input characteristic is the characteristic of a forward biased diode when V_{BE} is zero and I_B is also zero.

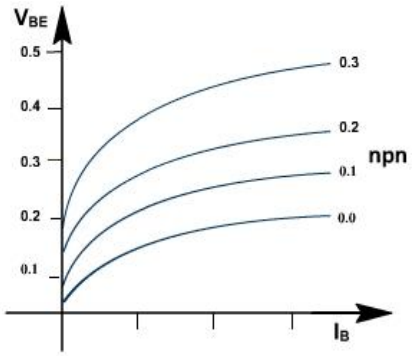


Fig.3.12

3.6.2 Output Characteristic:

The output characteristic is the curve between V_{CE} and I_C for various values of I_B. For fixed value of I_B and is shown in **fig. 3**.13. For fixed value of I_B, I_C is not varying much dependent on V_{CE} but slopes are greater than CE characteristic. The output characteristics can again be divided into three parts.

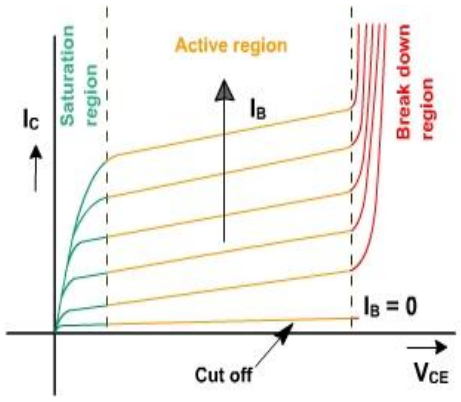


Fig. 3.13

(1) Active Region:

In this region collector junction is reverse biased and emitter junction is forward biased. It is the area to the right of V_{CE} = 0.5 V and above I_B= 0. In this region transistor current responds most sensitively to I_B. If transistor is to be used as an amplifier, it must operate in this region.

$$I_E = I_C + I_B$$

Since, $I_C = I_{Co} + \alpha_{dc} I_E$

$$I_C = I_{Co} + \alpha_{dc}(I_C + I_B)$$

or $(1 - \alpha_{dc}) I_C = \alpha_{dc} I_B + I_{CO}$

or $$I_C = \left(\frac{\alpha_{dc}}{1 - \alpha_{dc}}\right) I_B + \left(\frac{1}{1 - \alpha_{dc}}\right) I_{CO}$$

Let, $\beta_{dc} = \frac{\alpha_{dc}}{1 - \alpha_{dc}}$

$\therefore I_C = (1 + \beta_{dc}) I_{CO} + \beta_{dc} I_B$

β_{dc} is defined as current gain of the transistor is given by

$$\beta_{dc} = \frac{I_C - I_{CO}}{I_B + I_{CO}} \qquad \text{(E-19)}$$

If α_{dc} is truly constant then I_C would be independent of V_{CE}. But because of early effect, α_{dc} increases by 0.1% (0.001) e.g. from 0.995 to 0.996 as V_{CE} increases from a few volts to 10V. Then β_{dc} increases from 0.995 / (1-0.995) = 200 to 0.996 / (1-0.996) = 250 or about 25%. This shows that small change in α_{dc} reflects large change in β_{dc}. Therefore the curves are subjected to large variations for the same type of transistors.

(2) Cut Off:

Cut off in a transistor is given by $I_B = 0$, $I_C = I_{CO}$. A transistor is not at cut off if the base current is simply reduced to zero (open circuited) under this condition,

$$I_C = I_E = I_{CO} / (1 - \alpha_{dc}) = I_{CEO} \qquad \text{(E-20)}$$

The actual collector current with base open is designated as I_{CEO}. Since even in the neighborhood of cut off, α_{dc} may be as large as 0.9 for Ge, then $I_C = 10\ I_{CO}$(approximately), at zero base current. Accordingly in order to cut off transistor it is not enough to reduce I_B to zero, but it is necessary to reverse bias the emitter junction slightly. It is found that reverse voltage of 0.1 V is sufficient for cut off a transistor. In Si, the α_{dc} is very

nearly equal to zero, therefore, $I_C = I_{CO}$. Hence even with $I_B = 0$, $I_C = I_E = I_{CO}$ so that transistor is very close to cut off.

In summary, cut off means $I_E = 0$, $I_C = I_{CO}$, $I_B = -I_C = -I_{CO}$, and V_{BE} is a reverse voltage whose magnitude is of the order of 0.1 V for Ge and 0 V for Si.

3.6.3 Reverse Collector Saturation Current I_{CBO}:

When in a physical transistor emitter current is reduced to zero, then the collector current is known as I_{CBO} (approximately equal to I_{CO}). Reverse collector saturation current I_{CBO} also varies with temperature, avalanche multiplication and variability from sample to sample. Consider the circuit shown in **fig. 3.14**. V_{BB} is the reverse voltage applied to reduce the emitter current to zero.

$I_E = 0$, $I_B = -I_{CBO}$

If we require, $V_{BE} = -0.1$ V

Then $-V_{BB} + I_{CBO} R_B < -0.1$ V

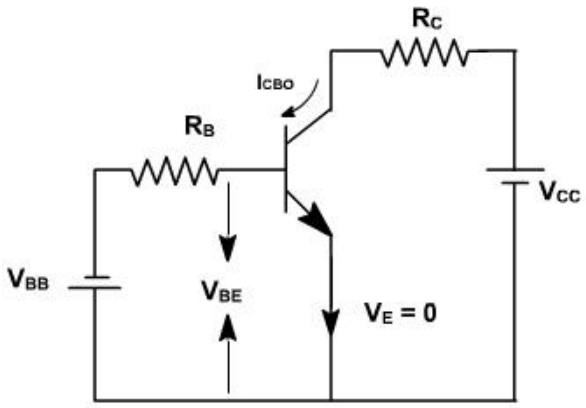


Fig-3.14

If $R_B = 100$ K, $I_{CBO} = 100$ m A, Then V_{BB} must be 10.1 Volts. Hence transistor must be capable to withstand this reverse voltage before breakdown voltage exceeds.

(3).Saturation Region:

In this region both the diodes are forward biased by at least cut in voltage. Since the voltage V_{BE} and V_{BC} across a forward is approximately 0.7 V therefore, $V_{CE} = V_{CB} + V_{BE} = - V_{BC} + V_{BE}$ is also few tenths of volts. Hence saturation region is very close to zero voltage axis, where all the current rapidly reduces to zero. In this region the transistor collector current is approximately given by V_{CC} / R_C and independent of base current. Normal transistor action is last and it acts like a small ohmic resistance.

3.6.4 Large Signal Current Gain β_{dc} :-

The ratio I_c / I_B is defined as transfer ratio or large signal current gain β_{dc}

$$\beta_{dc} = \frac{I_C}{I_B} \qquad \text{(E-21)}$$

Where I_C is the collector current and I_B is the base current. The β_{dc} is an indication if how well the transistor works. The typical value of β_{dc} varies from 50 to 300.

In terms of h parameters, β_{dc} is known as dc current gain and in designated h_{fE} ($\beta_{dc} = h_{fE}$). Knowing the maximum collector current and β_{dc} the minimum base current can be found which will be needed to saturate the transistor.

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

$$I_{B(min)} = \frac{I_{C(sat)}}{\beta_{dc}} \qquad \text{(E-22)}$$

This expression of β_{dc} is defined neglecting reverse leakage current (I_{CO}).

Taking reverse leakage current (I_{CO}) into account, the expression for the β_{dc} can be obtained as follows:

β_{dc} in terms of $_{dc}$ is given by

$$\beta_{dc} = \frac{\alpha_{dc}}{1-\alpha_{dc}}$$

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

$$= \frac{I_C - I_{CO}}{I_B + I_{CO}} \qquad \text{(E-23)}$$

Since, $I_{CO} = I_{CBO}$

$$\therefore \beta_{dc} = \frac{I_C - I_{CBO}}{I_B + I_{CBO}} \qquad \text{(E-24)}$$

Cut off of a transistor means $I_E = 0$, then $I_C = I_{CBO}$ and $I_B = -I_{CBO}$. Therefore, the above expression β_{dc} gives the collector current increment to the base current change form cut off to I_B and hence it represents the large signal current gain of all common emitter transistor.

3.9 Small Signal CE Amplifiers

CE amplifiers are very popular to amplify the small signal ac. After a transistor has been biased with a Q point near the middle of a dc load line, ac source can be coupled to the base. This produces fluctuations in the base current and hence in the collector current of the same shape and frequency. The output will be enlarged sine wave of same frequency.

The amplifier is called linear if it does not change the wave shape of the signal. As long as the input signal is small, the transistor will use only a small part of the load line and the operation will be linear.

On the other hand, if the input signal is too large. The fluctuations along the load line will drive the transistor into either saturation or cut off. This clips the peaks of the input and the amplifier is no longer linear.

The CE amplifier configuration is shown in **fig. 3.37**.

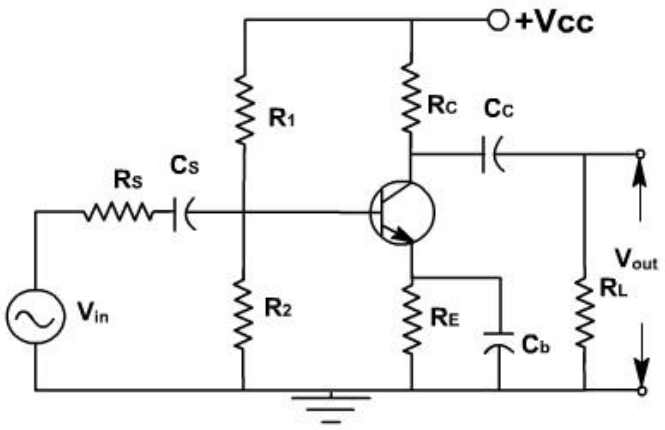


Fig. 3.37

The coupling capacitor (C_C) passes an ac signal from one point to another. At the same time it does not allow the dc to pass through it. Hence it is also called blocking capacitor.

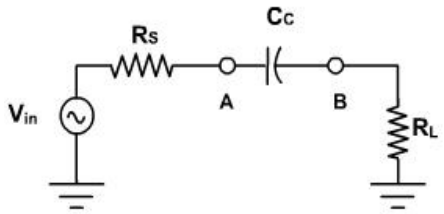


Fig. 3.38

For example in **fig. 3.38**, the ac voltage at point A is transmitted to point B. For this series reactance X_C should be very small compared to series resistance R_S. The circuit to the left of A may be a source and a series resistor or may be the Thevenin equivalent of a complex circuit. Similarly R_L may be the load resistance or equivalent resistance of a complex network. The current in the loop is given by

$$i = \frac{V_{in}}{\sqrt{(R_S + R_L)^2 + X_C^2}}$$

$$= \frac{V_{in}}{\sqrt{R^2 + X^2}} \qquad \text{(E-50)}$$

3.9.1 Analysis of CE amplifier:

In a transistor amplifier, the dc source sets up quiescent current and voltages. The ac source then produces fluctuations in these current and voltages. The simplest way to analyze this circuit is to split the analysis in two parts: dc analysis and ac analysis. One can use superposition theorem for analysis .

AC & DC Equivalent Circuits:

For dc equivalent circuit, reduce all ac voltage sources to zero and open all ac current sources and open all capacitors. With this reduced circuit shown in **fig. 3.3**9 dc current and voltages can be

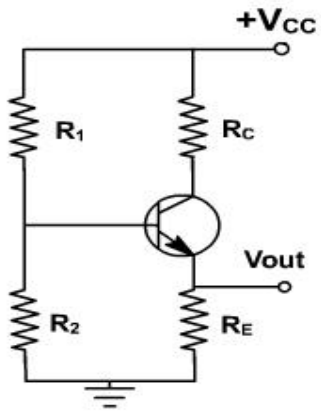


Fig-3.39

Calculated.

For ac equivalent circuits reduce dc voltage sources to zero and open current sources and short all capacitors. This circuit is used to calculate ac currents and voltage as shown in **fig.3.40**.

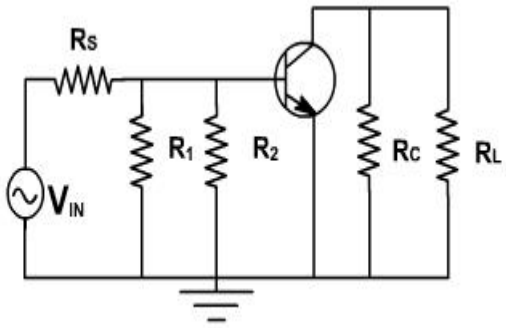


Fig.3.40

The total current in any branch is the sum of dc and ac currents through that branch. The total voltage across any branch is the sum of the dc voltage and ac voltage across that branch.

3.9.1.1 Phase Inversion:

Because of the fluctuation is base current; collector current and collector voltage also swings above and below the quiescent voltage. The ac output voltage is inverted with respect to the ac input voltage, meaning it is 180^{o} out of phase with input.

During the positive half cycle base current increase, causing the collector current to increase. This produces a large voltage drop across the collector resistor; therefore, the voltage output decreases and negative half cycle of output voltage is obtained. Conversely, on the negative half cycle of input voltage less collector current flows and the voltage drop across the collector resistor decreases, and hence collector voltage increases we get the positive half cycle of output voltage as shown in **fig.3.41**.

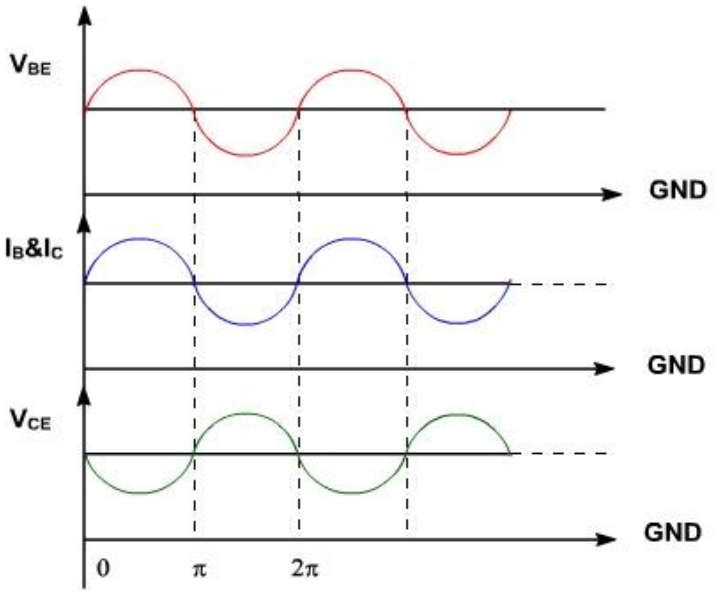


Fig.3.41

As frequency increases, $X_C\left(=\frac{1}{2\pi fc}\right)$ decreases, and current increases until it reaches to its maximum value v_{in} / R. Therefore the capacitor couples the signal properly from A to B when $X_C << R$. The size of the coupling capacitor depends upon the lowest frequency to be coupled. Normally, for lowest frequency $X_C \leq 0.1R$ is taken as design rule.

The coupling capacitor acts like a switch, which is open to dc and shorted for ac.

The bypass capacitor C_b is similar to a coupling capacitor, except that it couples an ungrounded point to a grounded point. The C_b capacitor looks like a short to an ac signal and therefore emitter is said ac grounded. A bypass capacitor does not disturb the dc voltage at emitter because it looks open to dc current.

3.9.1.2 AC Load line:

Consider the dc equivalent circuit **fig.3.42**.

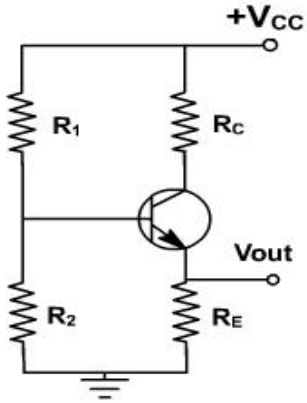


Fig.3.42

Assuming $I_C = I_C(approx)$, the output circuit voltage equation can be written as

$$V_{CE} = V_{CC} - I_C(R_C + R_E)$$

and $$I_C = -\frac{V_{CE}}{R_C + R_E} + \frac{V_{CC}}{R_C + R_E}$$

$$V_{CE} = 0, \quad I_C = \frac{V_{CC}}{R_C + R_E}$$

and $$I_C = 0, \quad V_{CE} = V_{CC}$$ (E-51)

The slop of the d.c load line is $-\frac{1}{R_C + R_E}$.

When considering the ac equivalent circuit, the output impedance becomes $R_C \parallel R_L$ which is less than $(R_C + R_E)$.

In the absence of ac signal, this load line passes through Q point. Therefore ac load line is a line of slope $(-1 / (R_C \parallel R_L))$ passing through Q point. Therefore, the output voltage fluctuations will now be corresponding to ac load line as shown in **fig. 3.43**. Under this condition, Q-point is not in the middle of load line, therefore Q-point is selected slightly upward, means slightly shifted to saturation side.

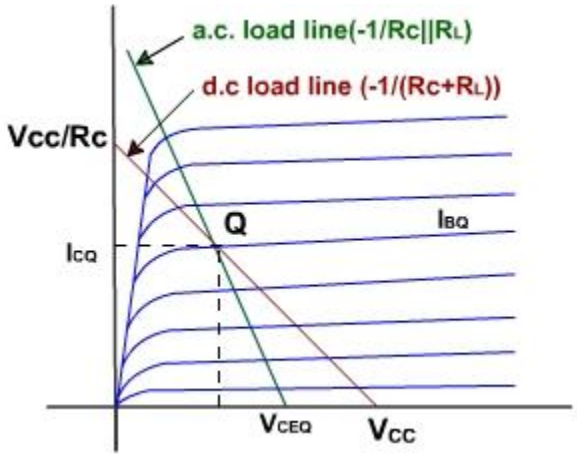


Fig.3.43

3.9.1.3 Voltage gain:

To find the voltage gain, consider an unloaded CE amplifier. The ac equivalent circuit is shown in **fig.3.44**. The transistor can be replaced by its collector equivalent model i.e. a current source and emitter diode which offers ac resistance r'_e.

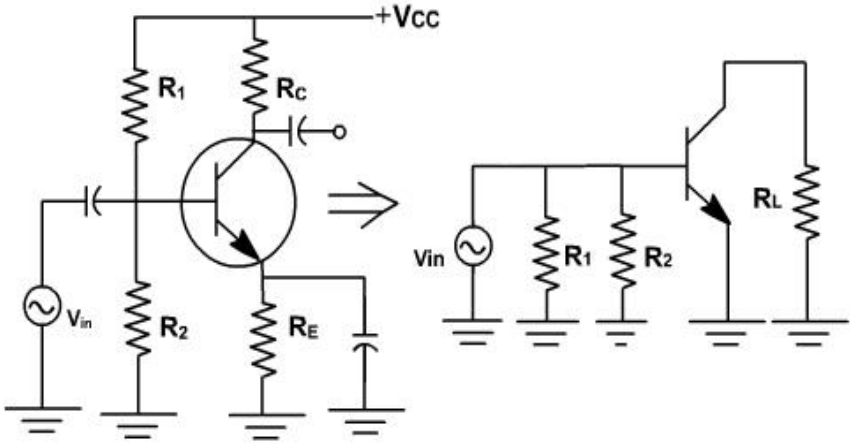


Fig. 3.44

The input voltage appears directly across the emitter diode.

Therefore emitter current $i_e = V_{in} / r'_e$.

Since, collector current approximately equals emitter current and $i_C = i_e$ and $v_{out} = - i_e R_C$ (The minus sign is used here to indicate phase inversion)

Further $v_{out} = - (V_{in} R_C) / r'_e$

Therefore voltage gain $A = v_{out} / v_{in} = -R_C / r'_e$

The ac source driving an amplifier has to supply alternating current to the amplifier. The input impedance of an amplifier determines how much current the amplifier takes from the ac source.

In a normal frequency range of an amplifier, where all capacitors look like ac shorts and other reactance are negligible, the ac input impedance is defined as

$z_{in} = v_{in} / i_{in}$

Where v_{in}, i_{in} are peak to peak values or rms values

The impedance looking directly into the base is symbolized $z_{in\,(base)}$ and is given by

$Z_{in(base)} = v_{in} / i_b$,

Since, $v_{in} = i_e r'_e$

$\approx \beta_i \beta r'_e$

$z_{in\,(base)} = \beta r'_e$.

From the ac equivalent circuit, the input impedance z_{in} is the parallel combination of R_1 , R_2 and $\beta r'_e$.

$Z_{in} = R_1 \| R_2 \| \beta r'_e$

The Thevenin voltage appearing at the output is

$v_{out} = A\, v_{in}$

The Thevenin impedance is the parallel combination of R_C and the internal impedance of the current source. The collector current source is an ideal source, therefore it has an infinite internal impedance.

$z_{out} = R_C$.

The simplified ac equivalent circuit is shown in **fig.3.45**.

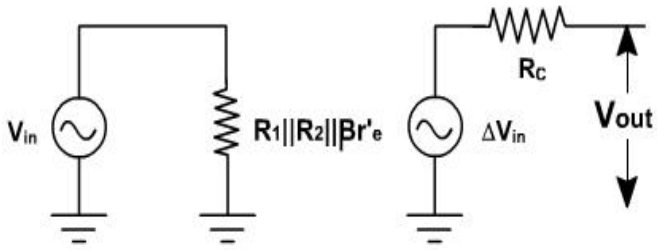


Fig. 3.45

Example-1:

Select R_1 and R_2 for maximum output voltage swing in the circuit shown in **fig.3.46**.

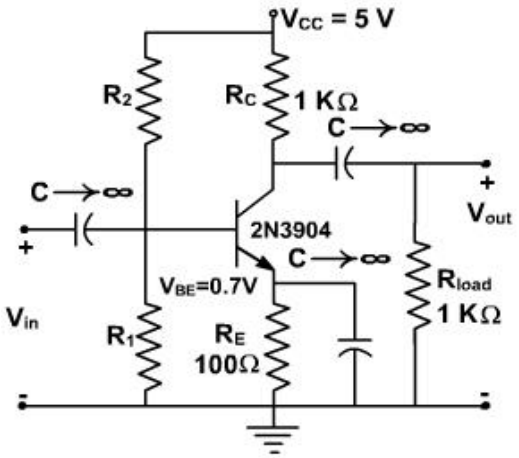


Fig.3.46

Solution:

We first determine I_{CQ} for the circuit

$$R_{ac} = R_C \parallel R_{load} = 500$$
$$R_{dc} = R_E + R_C = 1100$$
$$I_{CQ} = \frac{V_{CC}}{R_{ac} + R_{dc}} = \frac{5}{500 + 1100} = 3.13 \text{ mA}$$

For maximum swing,

$V'_{CC} = 2\ V_{CEQ}$

The quiescent value for V_{CE} is the given by

V_{CEQ}= (3.13 mA) (500 Ω) = 1.56 V

The intersection of the ac load line on the v_{CE} axis is V'_{CC} = 3.13V. From the manufacturer's specification, β for the 2N3904 is 180. R_B is set equal to 0.1 βR_E. So,

R_B = 0.1(180)(100) = 1.8 K□ Ω□

V_{BB} = (3.13 x 10-3) (1.1 x 100) + 0.7 = 1.044 V

Since we know V_{BB} and R_B, we find R_1 and R_2,

$$R_1 = \frac{R_B}{1 - V_{BB} / V_{CC}} = \frac{1800}{1 - 1.044 / 5} = 2.28 \text{ K}\Omega$$
$$R_2 = \frac{R_B V_{CC}}{V_{BB}} = \frac{1800 \times 5}{1.044} = 8.62 \text{ K}\Omega$$

The maximum output voltage swing, ignoring the non-linearity's at saturation and cutoff, would then be

$$\text{Max collector current swing} = 2\ I_{CQ} (R_C \parallel R_{load})$$
$$= 2 (3.13 \text{ mA}) (500\ \Omega) = 3.13 \text{ V}$$

The load lines are shown on the characteristics of **fig.3.47**.

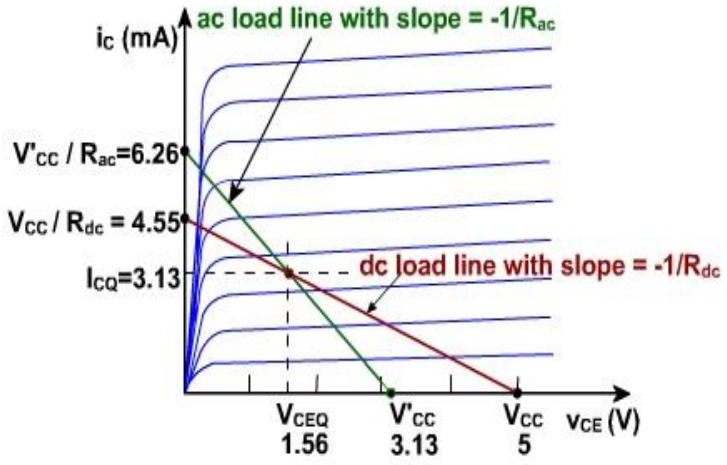


Fig.3.47

The maximum power dissipated by the transistor is calculated to assure that it does not exceed the specifications. The maximum average power dissipated in the transistor is

$P_{(transistor)} = V_{CEQ}\, I_{CQ}$ = (1.56 (V)) (3.13 mA) =4.87 mW

This is well within the 350 mW maximum given on the specification sheet. The maximum conversion efficiency is

$$\eta = \frac{P_{out}(ac)}{P_{VCC}(dc)} = \frac{\left(3.13 \times 10^{-3}/2\right)^2 \times 1000/2 \times 100}{5 \times 3.13 \times 10^{-3} + 5^2/10.9 \times 10^3} = 6.84\%$$

3.10 Design of Amplifier

Example -1 (Common Emitter Amplifier Design)

Design a common-emitter amplifier with a transistor having a β =200 and V_{BE} = 0.7 V. Obtain an overall gain of |A V | ≥ 100 and maximum output voltage swing. Use the CE configuration shown in **fig.3.48** with two power supplies. R_{source} is the resistance associated with the source, v_{source}. Let R_{source}= 100 Ohms. The output

load is 2KΩ. Determine the resistor values of the bias circuitry, the maximum undistorted output voltage swing, and the stage voltage gain.

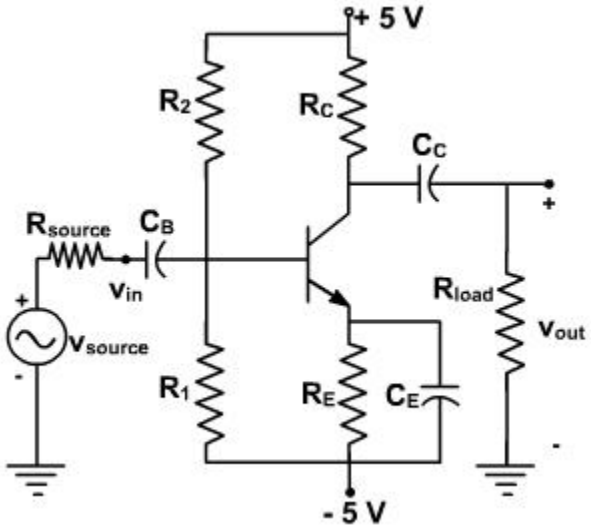


Fig. 3.48

Solution:

The maximum voltage across the amplifier is 10 V since the power supply can be visualized as a 10V power supply with a ground in the center. In this case, the ground has no significance to the operation of the amplifier since the input and output are isolated from the power supplies by capacitors.

We will have to select the value for R_C and we are really not given enough information to do so. Let choose $R_C = R_{load}$.

We don't have enough information to solve for R_B – we can't use the bias stability criterion since we don't have the value of R_E either. We will have to (arbitrarily) select a value of R_B or R_E. If this leads to a contradiction, or "bad" component values (e.g., unobtainable resistor values), we can come back and modify our choice. Let us select a value for R_E that is large enough to obtain a reasonable value of V_{BB}, Selecting R_E as 400Ω will not appreciably reduce the collector current yet it will help in maintaining a reasonable value of V_{BB}. Thus,

$R_B = 0.1\ \beta\ R_E = 0.1\ (200)(400) = 8\ K\ \Omega$

To insure that we have the maximum voltage swing at the output, we will use

$$I_{CQ} = \frac{V_{CC}}{R_{ac} + R_{dc}} = \frac{10}{1000 + 2400} = 2.94\ mA$$

$$V_{BB} = V_{BE} + I_{CQ}\ (R_B/\beta + R_E) = 0.7 + 2.9 \times 10^{-3}\left(\frac{8000}{200} + 400\right) = 1.99\ V$$

Note that we are carrying out our calculations to four places so that we can get accuracy to three places. The bias resistors are determined by

$$R_1 = \frac{R_B}{1 - V_{BB}/V_{CC}} = \frac{8000}{1 - 1.99/10} = 9.99\ K\Omega$$

$$R_2 = R_B \frac{V_{CC}}{V_{BB}} = 8000\left(\frac{10}{1.99}\right) = 40.2\ K\Omega$$

Since we designed the bias circuit to place the quiescent point in the middle of the ac load line, we can use

V_{out}(undistorted p-p) $1.8\ (2.94 \times 10^{-3})\ (2\ K\ \Omega \parallel 2\ K\ \Omega) = 5.29\ V$

Now we can determine the gain of the amplifier itself.

$$|A_v| = g_m\ (R_C \parallel R_{load}) = \frac{2.94 \times 10^{-3} \times 1000}{26 \times 10^{-3}}$$

Using voltage division, we can determine the gain of the overall circuit.

The value of Rin can be obtained as

$$R_{in} = r_\pi \parallel R_B = 1.77k\Omega \parallel 8k\Omega = 1.45\ k\Omega$$

Thus the overall gain of the amplifier is

$$|A_v|\ \text{overall} = \left|\frac{v_{out}}{v_{in}} = 113\right| x \frac{R_{in}}{R_{in} + R_{source}} = 106$$

This shows that the common-emitter amplifier provides high voltage gain. However, it is very noisy, it has a low input impedance, and it does not have the stability of the emitter resistor common emitter amplifier.

Example-2 (Emitter-Resistor Amplifier Design)

Design an emitter-resistor amplifier as shown in **fig.3.49** to drive a 2 KΩ load using a pnp silicon transistor, V_{CC} = -24V, β = 200, A_v = -10, and V_{BE} = -0.7 V. Determine all element values and calculate A_i, R_{in}, I_{CQ} and the maximum undistorted symmetrical output voltage swing for three values of R_C as given below:

1. $R_C = R_{load}$
2. $R_C = 0.1\ R_{load}$
3. $R_C = 10\ R_{load}$

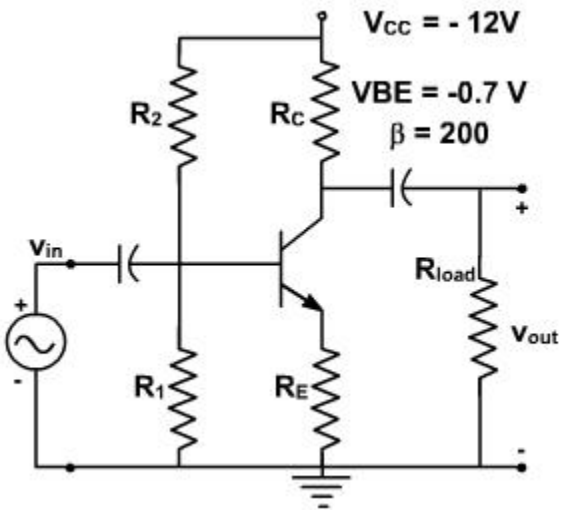


Fig. 3.49

Solution:

(a) $R_C = R_{load}$

We use the various equations derived in previous lecture in order to derive the parameters of the circuit.

From the voltage gain, we can solve for R'_E.

$$A_v = -10 = -\frac{R_{load} \,||\, R_C}{r_e + R_E} = -\frac{2K\Omega \,||\, 2K\Omega}{r_e + R_E}$$

So $R'_E = r_e + R_E = 100\ \Omega$

We can find the quiescent value of the collector current I_C form the collector-emitter loop using the equation for the condition of maximum output swing.

$$I_{CQ} = \frac{V_{CC}}{R_{dc} + R_{ac}} = -7.5\ mA$$

Therefore, $r_E' = \frac{25 \times 10^{-3}}{7.5 \times 10^{-3}} = 3.33\Omega$

This is small enough that we shall ignore it to find that $R_E = 100\ \Omega$. Since we now know β and R_E. We can use the design guideline.

$R_B = 0.1\ \beta\ R_E = 2\ k\ \Omega$

As designed earlier, the biasing circuitry can be designed in the same manner and given by

$V_{BB} = -1.52\ V$

$R_1 = 2.14\ K\ \Omega$

$R_2 = 3.6\ K\ \Omega$

The maximum undistorted symmetrical peak to peak output swing is then

V_{out} (P-P) = 1.8 I_{CQ} (R_{load} || R_C) = 13.5 V

Thus current gain A_i = -9.1

and input impedance R_{in} = 1.82 K Ω

(b) $R_C = 0.1\ R_{load}$

we repeat the steps of parts (a) to find

R_C =200 Ω R_i = 390 Ω

I_{CQ} =-57.4 mA R_2 =4.7K Ω

r'_e = 0.45 Ω v_{out}(p-p) = 18.7 V

R_B = 360 Ω A_i = -1.64

V_{BB} = -1.84 V R_{in} = 327 Ω

(C) R_C =10 R_{load}

Once again, we follow the steps of part (a) to find

R_C =20 K Ω R_1 =3.28K Ω

I_{CQ} =-1.07 mA R_2 = 85.6K Ω

r'_e = 24.2 Ω v_{out}(p-p) = 3.9 V

R_B = 3.64K Ω A_i =-14.5

V_{BB} = -0.886 V R_{in} = 2.91K W

We now compare the results obtained Table-I for the purpose of making the best choice for R_C.

	I_{CQ}	**A_i**	**R_{in}**	**v_{out}(p-p)**

$R_C = R_{load}$	-7.5 mA	-9.1	1.82K W	13.5 V
$R_C = 0.1\ R_{load}$	-57.4 mA	-1.64	327 W	20.8 V
$R_C = 10\ R_{load}$	-1.07mA	-14.5	2.91W	3.9 V

Table -3.1 Comparsion for the three selections of R_C

It indicates that of the three given ratios of R_C to R_{load}, $R_C = R_{load}$ has the most desirable performance in the CE amplifier stage.

It can be used as a guide to develop a reasonable designs. In most cases, this choice will provide performance that meets specifications. In some applications, it may be necessary to do additional analysis to find the optimum ratio of R_C to R_{load}.

Example- 3 (Capacitor-Coupled Emitter-Resistor Amplifier Design)

Design an emitter-resistor amplifier as shown in **fig. 3.50** with A_V =-10, β =200 and R load = 1K Ω. A pnp transistor is used and maximum symmetrical output swing is required.

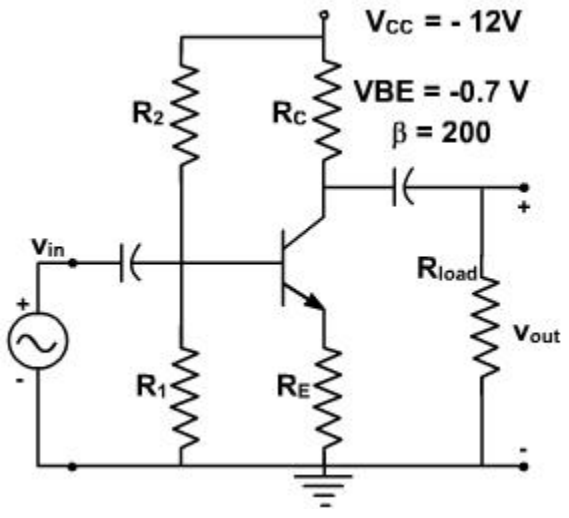


Fig. 3.50

Solution:

As designed earlier, we shall chose $R_C = R_{load} = 10\ k\Omega$.

The voltage gain is given by $A_v = \frac{R_{load} \parallel R_C}{R'_E}$

where $R'_E = R_E + r'_e$.

Substituting A_V, R_{load} and R_C in this equation, we find $R'_E = 50\ \Omega$.

We need to know the value of r'_e to fine R_E. We first find R_{ac} and R_{dc}, and then calculate the Q point as follows (we assume r'_e is small, so $R_E = R'_E$)

$R_{ac} = R_E + R_C \parallel R_{load} = 550\ \Omega$

$R_{dc} = R_E + R_C = 1050\ \Omega$

Now, the first step is to calculate the quiescent collector current needed to place the Q-point into the center of the ac load line (i.e., maximum swing). The equation is

$$I_{CQ} = \frac{V_{CC}}{R_{ac} + R_{dc}} = -7.5\ mA$$

The quantity, r'_e , is found as follows

$$r_e' = \frac{25(mV)}{|I_{CQ}|} = \frac{25(mV)}{7.5(mA)} = 3.33\Omega$$

Then

$R_E = 50 - r_e = 46.67\ \Omega$

If there were a current gain or input resistance specification for this design, we would use it to solve for the value of R_B. Since is no such specification, we use the expression

$R_B = 0.1\ \beta\ R_E = 0.1\ (200)\ (46.6) = 932\ \Omega$

Then continuing with the design steps,

$$A_i = \frac{-R_B}{R_B/\beta + r_E' + R_B} \cdot \frac{R_C}{R_C + R_{load}} = -8.50$$

$$V_{CEQ} = V_{CC} - (R_C + R'_E)\, I_{CQ} = -4.125\ V$$

and

$$V_{BB} = I_{CQ} - \left(R_E + \frac{R_E}{\beta}\right) + V_{BE} = -1.08\ V$$

$$R_1 = \frac{R_B}{1 - V_{BB}/V_{CC}} = 1.02\ K\Omega$$

$$R_2 = \frac{R_B V_{CC}}{V_{BB}} = 10.3\ K\Omega$$

$$R_{in} = \frac{R_B(r_e + R_E)}{R_B/\beta + r_e + R_E} = 8.51\Omega$$

$$R_O = R_C = 1\ K\Omega$$

The last equality assumes that r_O is large compared to R_C.

The maximum undistorted peak to peak output swing is given by

$1.8\ |\ I_{CQ}\ |\ (\ R_C\ ||\ R_{load}\ ) = 1.8\ (\ 0.0075\ )\ (\ 500\ ) = 6.75\ V$

The power delivered into the load and the maximum power dissipated by the transistor are found as

$$P_{Load} = \frac{1}{2}\left(I_{CQ}\frac{R_C}{R_C + R_{load}}\right)^2 R_{load} = \frac{I_{CQ}^2 R_{load}}{8} = 7\ \ mW$$

$$P_{transistor} = V_{CEQ} I_{CQ} = (-4.125V)(-7.5mA) = 31\,mW$$

The load lines for this circuit are shown in **fig. 3.51**.

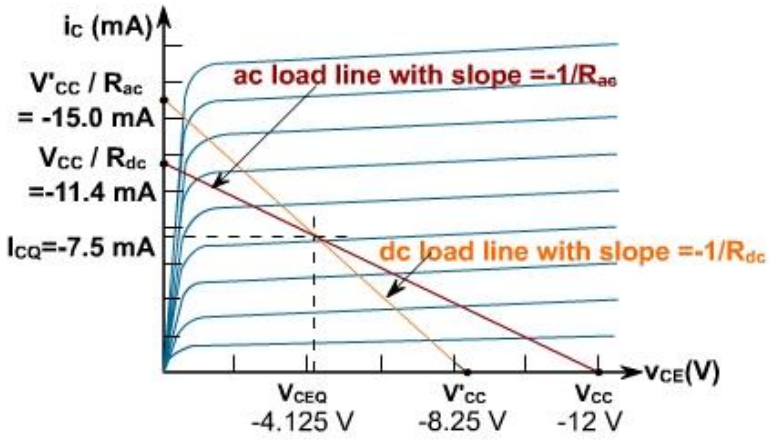


Fig. 3.51

3.11 Common Collector Amplifier:

If a high impedance source is connected to low impedance amplifier then most of the signal is dropped across the internal impedance of the source. To avoid this problem common collector amplifier is used in between source and CE amplifier. It increases the input impedence of the CE amplifier without significant change in input voltage.

Fig.3.52 shows a common collector (CC) amplifier. Since there is no resistance in collector circuit, therefore collector is ac grounded. It is also called grounded collector amplifier. When input source drives the base, output appears across emitter resistor. A CC amplifier is like a heavily swamped CE amplifier with a collector resistor shorted and output taken across emitter resistor.

$v_{out} = v_{in} - v_{BE}$

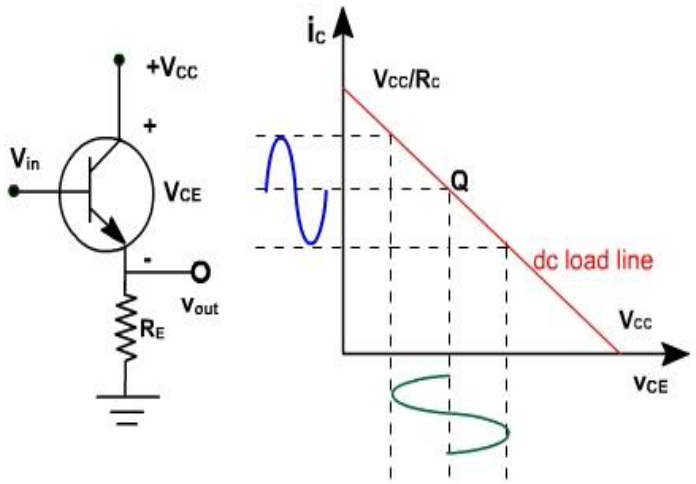


Fig. 3.52

Therefore, this circuit is also called emitter follower, because V_{BE} is very small. As v_{in} increases, v_{out} increases.

If v_{in} is 2V, v_{out} = 1.3V

If v_{in} is 3V, v_{out} = 2.3V.

Since v_{out} follows exactly the v_{in} therefore, there is no phase inversion between input and output.

The output circuit voltage equation is given by

$$V_{CE} = V_{CC} - I_E R_E \qquad (3.52)$$

Since $I_E \approx I_C$

$$\therefore I_C = (V_{CC} - V_{CE}) / R_E \qquad (3.53)$$

This is the equation of dc load line. The dc load line is shown in **Fig.3.5**2.

3.11.1 Voltage gain:

Fig.3.53, shows an emitter follower driven by a small ac voltage**.** The input is applied at the base of transistor and output is taken across the emitter resistor. **Fig.3.54**, shows the ac equivalent circuit of the amplifier. The emitter is replaced by ac resistance r'_e.

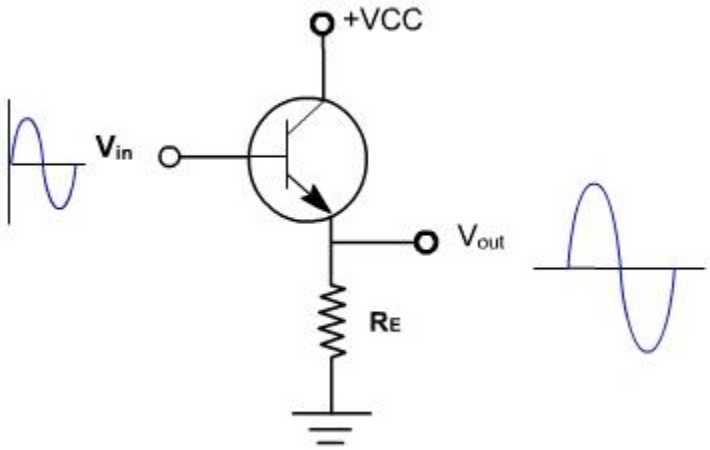


Fig. 3.53

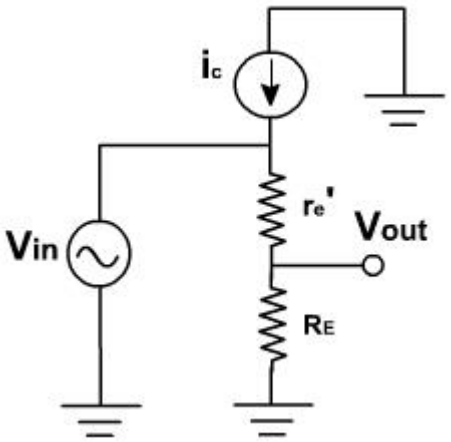


Fig. 3.54

The ac output voltage is given by

$v_{out} = R_E\, i_e$

and, $v_{in} = i_e\,(R_E + r'_e)$

Therefore, $A = R_E / (R_E + r'_e)$

Since $r'_e << R_E$

□ □ A □ □ 1.

Therefore, it is a unity gain amplifier. The practical emitter follower circuit is shown in **Fig.3.55**.

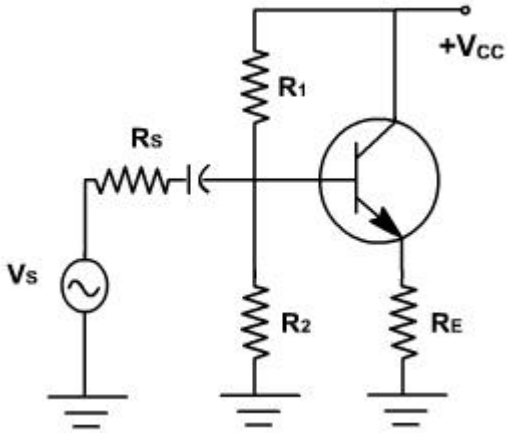


Fig. 3.55

The ac source (v_S) with a series resistance R_S drives the transistor base. Because of the biasing resistor and input impedance of the base, some of the ac signal is lost across the source resistor. The ac equivalent circuit is shown in **Fig. 3.5**6.

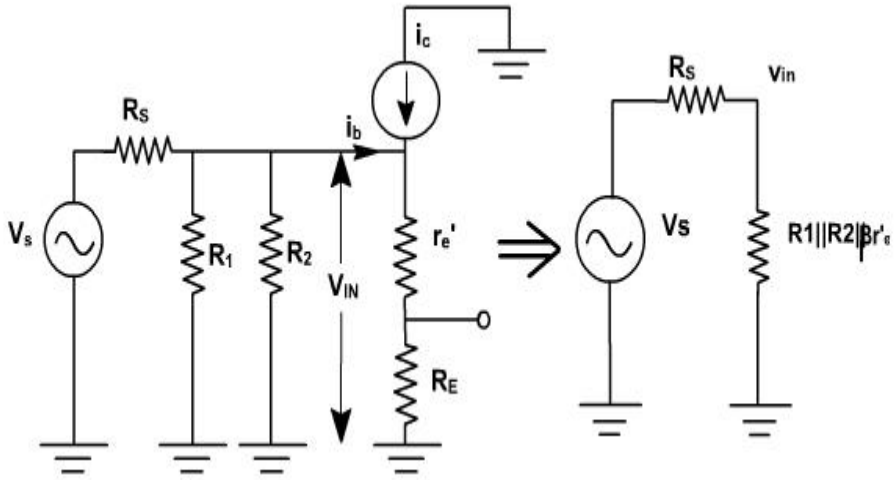


Fig.3.56

The input impedance at the base is given by

$$Z_{in(base)} = \frac{v_{in}}{i_B}$$
$$= \frac{i_e(r'_e + R_E)}{i_b}$$
$$= \frac{\beta i_b (r'_e + R_E)}{i_b}$$
$$= \beta(r'_e + R_E)$$

Since r'_e is very small in comparison with R_E

$$\therefore Z_{in(base)} \approx \beta R_E \qquad (3.54)$$

The total input impedance of an emitter follower includes biasing resistors in parallel with input impedance of the base.

$$z_{in} = R_1 \parallel R_2 \parallel \beta (r'_e + R_E) \qquad (3.55)$$

Since βR_E is very large as compared to R_1 and R_2.

Thus, $z_{in} \approx R_1 \parallel R_2$ (3.56)

Therefore input impedance is very high.

Applying Thevenin's theorem to the base circuit of **Fig. 3.56**, it becomes a source v_{in} and a series resistance ($R_1 \parallel R_2 \parallel R_S$) as shown in **Fig.3.57**.

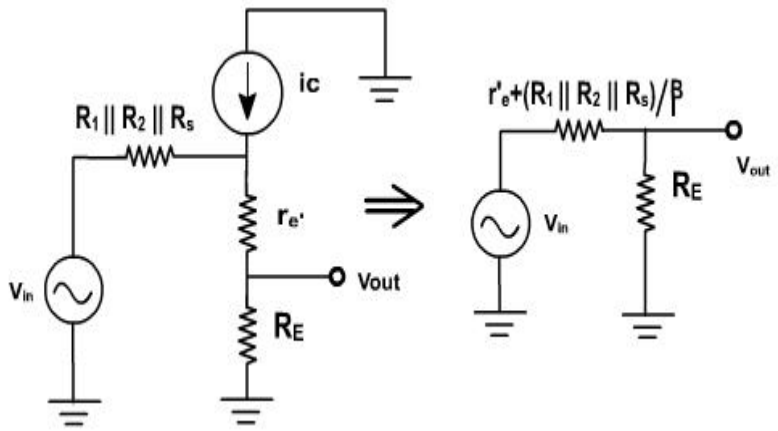


Fig. 3.57

$$v_{in} = (R_1 \| R_2 \| R_S) + i_e\ (r'_e + R_E)$$

$$\text{or,} \quad i_e = \frac{V_{in}}{R_E + r'_e + \frac{R_1 \| R_2 \| R_S}{\beta}}$$

The emitter resistor R_E is driven by an ac source with output impedance of

$$Z_{out} = r'_e + \frac{R_1 \| R_2 \| R_S}{\beta}$$

The impedance of the amplifier seen from the output terminal is given by

$$Z = R_E \| \ r'_e + \frac{R_1 \| R_2 \| R_S}{\beta}$$

The ouput voltage is given by

$$V_{out} = A v_{in}$$

$$= \frac{R_E}{R_E + r'_e + \frac{R_1 \| R_2 \| R_S}{\beta}} v_{in}$$

$$\approx v_{in} \quad (\text{if } R_E \text{ is very larage}) \qquad (3.57)$$

Example 1:

Find the Q-point of the emitter follower circuit of **fig.3.58** with R_1 = 10 KΩ and R_2 = 20 KΩ. Assume the transistor has a β of 100 and input capacitor C is very-very large.

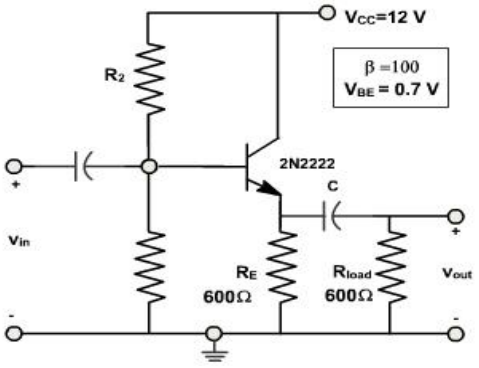


Fig. 3.58

Solution:

We first find the Thevenin's equivalent of the base bias circuitry.

$R_B = R_1 \parallel R_2 = 6.67\ K\Omega$

$$V_{BB} = \frac{R_1 V_{CC}}{R_1 + R_2} = \frac{12(10^4)}{30 \times 10^3} = 4\ V$$

From the bias equation we have

$$I_C = I_{CQ} = \frac{V_{BB} - V_{BE}}{\frac{R_B}{\beta} + R_E} = \frac{4 - 0.7}{\frac{6670}{100} + 600} = 4.95\ mA$$

Example - 2

Find the output voltage swing of the circuit of **fig.3.5**8.

Solution:

The Q-Point location has already been calculated in **Example-1**. We found that the quiescent collector current is 4.95 mA.

The Output voltage swing = 2 **.** I_C peak **.** ($R_E \parallel R_{Load}$) = $2(4.95 \times 10^{-3})$ (300) = 2.97V

This is less than the maximum possible output swing. Continuing the analysis,

$V_{CEQ} = V_{CC} - I_{CQ} R_E = 9.03\ V$

$V'_{CC} = V_{CEQ} + I_{CQ} (R_E \parallel R_{Load}) = 10.5\ V$

$$I'_{CC} = \frac{10.5}{300} = 35.1\ mA$$

The load lines for this problem are shown in **Fig.3.59**.

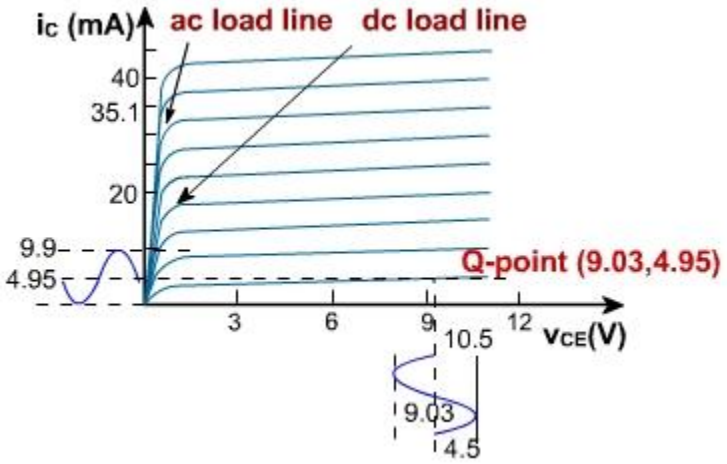


Fig.3.59

3.12 Small signal low frequency transistor Models:

All the transistor amplifiers are two port networks having two voltages and two currents. The positive directions of voltages and currents are shown in **fig.3.60**.

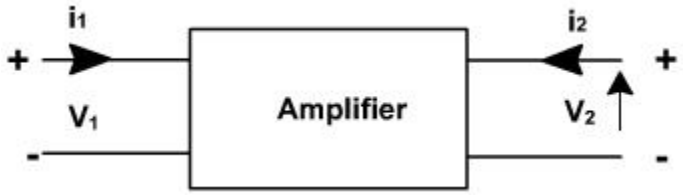


Fig.3.60

Out of four quantities two are independent and two are dependent. If the input current i_1 and output voltage v_2 are taken independent then other two quantities i_2 and v_1 can be expressed in terms of i_1 and V_2.

$$v_1 = f_1(i_1, v_2)$$
$$i_2 = f_2(i_1, v_2) \qquad \text{(E-3.58)}$$

The equations can be written as

$$v_1 = h_{11} i_1 + h_{12} v_2$$
$$i_2 = h_{21} i_1 + h_{22} v_2 \quad (\text{E-3.59})$$

where h_{11}, h_{12}, h_{21} and h_{22} are called h-parameters.

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

= h_i = input impedance with output short circuit to ac.

$$h_{12} = \left.\frac{v_1}{v_2}\right|_{i_2 = 0}$$

=h_r = fraction of output voltage at input with input open circuited or reverse voltage gain with input open circuited to ac (dimensions).

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

= h_f = negative of current gain with output short circuited to ac.

The current entering the load is negative of I_2. This is also known as forward short circuit current gain.

$$h_{22} = \left.\frac{i_2}{i_1}\right|_{i_2 = 0}$$

= h_o = output admittance with input open circuited to ac.

If these parameters are specified for a particular configuration, then suffixes e,b or c are also included, e.g. h_{fe} ,h_{ib} are h parameters of common emitter and common collector amplifiers

Using two equations the generalized model of the amplifier can be drawn as shown in **fig. 3.61**

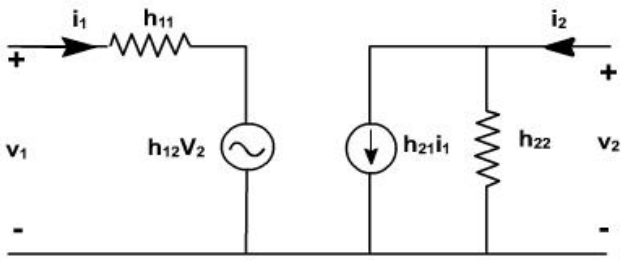


Fig. 3.61

3.21.1h-Parameters

The hybrid model for a transistor amplifier can be derived as follow:

Let us consider CE configuration as show in **fig. 3.62**. The variables, i_B, i_C ,v_C, and v_B represent total instantaneous currents and voltages i_B and v_C can be taken as independent variables and v_B, I_C as dependent variables.

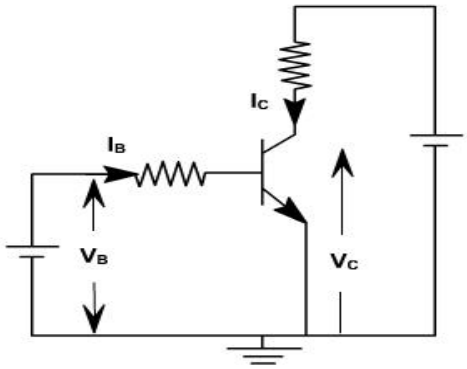


Fig. 3.62

$v_B = f_1 (i_B , v_C)$

$I_C = f_2 (i_B , v_C)$.

Using Taylor 's series expression, and neglecting higher order terms we obtain.

$$\Delta v_B = \left.\frac{\partial f_1}{\partial i_B}\right|_{v_C} \Delta i_B + \left.\frac{\partial f_1}{\partial v_C}\right|_{i_B} \Delta v_C$$

$$\Delta i_C = \left.\frac{\partial f_2}{\partial i_B}\right|_{v_C} \Delta i_B + \left.\frac{\partial f_2}{\partial v_C}\right|_{i_B} \Delta v_C$$

The partial derivatives are taken keeping the collector voltage or base current constant. The Δv_B, Δv_C, Δi_B, Δi_C represent the small signal (incremental) base and collector current and voltage and can be represented as v_b ,i_b ,v_C ,i_C.

$$\therefore v_b = h_{ie}\, i_B + h_{re}\, v_C$$
$$i_C = h_{fe}\, i_B + h_{oe}\, v_b$$

where

$$h_{ie} = \left.\frac{\partial f_1}{\partial i_B}\right|_{v_C} = \left.\frac{\partial v_B}{\partial i_B}\right|_{v_C}; \qquad h_{re} = \left.\frac{\partial f_1}{\partial v_C}\right|_{i_B} = \left.\frac{\partial v_B}{\partial v_C}\right|_{i_B}$$

$$h_{fe} = \left.\frac{\partial f_2}{\partial i_B}\right|_{v_C} = \left.\frac{\partial i_C}{\partial i_B}\right|_{v_C}; \qquad h_{oe} = \left.\frac{\partial f_2}{\partial v_C}\right|_{i_B} = \left.\frac{\partial v_B}{\partial v_C}\right|_{i_B} \qquad \text{(E-3.60)}$$

The model for CE configuration is shown in **fig. 3.63**.

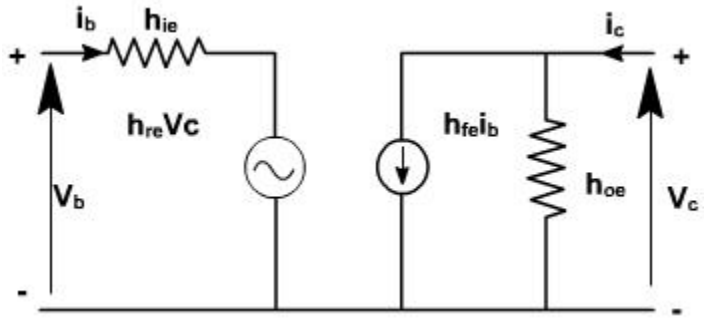


Fig. 3.63

3.12.2 Determination of h - parameters:

To determine the four h-parameters of transister amplifier, input and output characteristic are used. Input characteristic depicts the relationship between input voltage and input current with output voltage as parameter. The output characteristic depicts the relationship between output voltage and output current with input current as parameter. **Fig.3.64**, shows the output characterisitcs of CE amplifier.

$$h_{fe} = \left.\frac{\partial i_C}{\partial i_B}\right|_{V_C} = \frac{i_{C2} - i_{C1}}{i_{b2} - i_{b1}}$$

(E-3.61)

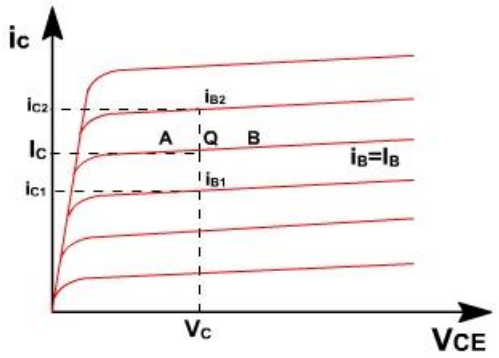


Fig.3.64

The current increments are taken around the quiescent point Q which corresponds to $i_B = I_B$ and to the collector voltage $V_{CE} = V_C$

$$h_{oe} = \left.\frac{\partial i_C}{\partial V_C}\right|_{i_B}$$

(E-3.62)

The value of h_{oe} at the quiescent operating point is given by the slope of the output characteristic at the operating point (i.e. slope of tangent AB).

$$h_{ie} = \frac{\partial V_B}{\partial i_B} \approx \left.\frac{\Delta V_B}{\Delta i_B}\right|_{V_C}$$

(E-3.63)

h_{ie} is the slope of the appropriate input on **fig. 3.65**, at the operating point (slope of tangent EF at Q).

$$h_{re} = \frac{\partial V_B}{\partial V_C} = \frac{\Delta V_B}{\Delta V_C}\bigg|_{I_B} = \frac{V_{B2} - V_{B1}}{V_{C2} - V_{C1}}$$

(E-3.64)

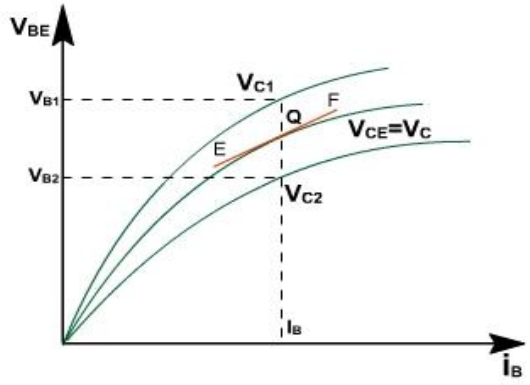


Fig. 3.65

A vertical line on the input characteristic represents constant base current. The parameter hre can be obtained from the ratio (V_{B2}– V_{B1}) and (V_{C2}– V_{C1}) for at Q.

Typical CE h-parametersof transistor 2N1573 are given below:

h_{ie}=1000 ohm.

h_{re}=2.5*10–4

h_{fe}= 50

h_{oe} = 25 μ□ A/ V

3.12.3 Analysis of a transistor amplifier using h-parameters:

To form a transistor amplifier it is only necessary to connect an external load and signal source as indicated in **fig.3.66** and to bias the transistor properly.

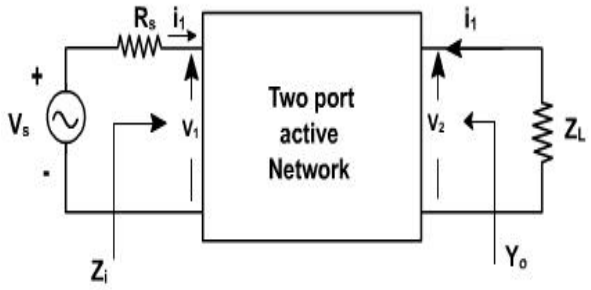


Fig.3.66

Consider the two-port network of CE amplifier. R_S is the source resistance and Z_L is the load impedence h-parameters are assumed to be constant over the operating range. The ac equivalent circuit is shown in **fig.3.67**. (Phasor notations are used assuming sinusoidal voltage input). The quantities of interest are the current gain, input impedence, voltage gain, and output impedence.

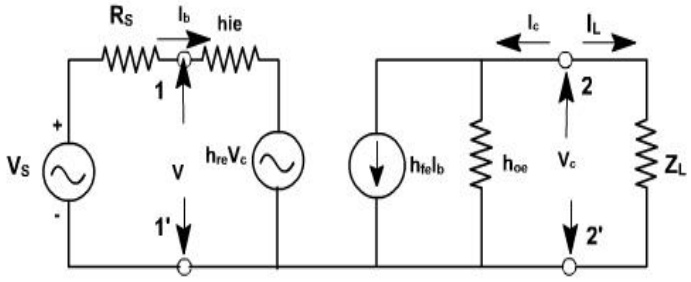


Fig.3.67

Current gain:

For the transistor amplifier stage, A_i is defined as the ratio of output to input currents.

$$A_i = \frac{I_L}{I_b} = \frac{-I_C}{I_b} \qquad (I_L + I_c = 0. \quad \therefore I_L = -I_c)$$

$$I_C = h_{fe} I_b + h_{oe} V_c$$

$$V_c = I_L Z_L = -I_c Z_L$$

$$\therefore I_c = h_{fe} I_b + h_{oe} (-I_c Z_L)$$

$$\text{or } \frac{I_c}{I_b} = \frac{h_{fe}}{1 + h_{oe} Z_L}$$

$$\therefore A_i = - \frac{h_{fe}}{1 + h_{oe} Z_L}$$

(E-3.65)

Input Impedence:

The impedence looking into the amplifier input terminals (1,1') is the input impedence Z_i

$$Z_i = \frac{V_b}{I_b}$$

$$V_b = h_{ie} I_b + h_{re} V_c$$

$$\frac{V_b}{I_b} = h_{ie} + h_{re} \frac{V_c}{I_b}$$

$$= h_{ie} - \frac{h_{re} I_c Z_L}{I_b}$$

$$\therefore Z_i = h_{ie} + h_{re} A_i Z_L$$

$$= h_{ie} - \frac{h_{re} h_{fe} Z_L}{1 + h_{oe} Z_L}$$

$$\therefore Z_i = h_{ie} - \frac{h_{re} h_{fe}}{Y_L + h_{oe}} \qquad (\text{since } Y_L = \frac{1}{Z_L})$$

(E-3.66)

Voltage gain:

The ratio of output voltage to input voltage gives the gain of the transistors.

$$A_v = \frac{V_C}{V_b} = -\frac{I_C Z_L}{V_b}$$

$$\therefore A_v = \frac{I_B A_i Z_L}{V_b} = \frac{A_i Z_L}{Z_i} \qquad \text{(E-3.67)}$$

Output Admittance:

It is defined as

$$Y_O = \left.\frac{I_c}{V_c}\right|_{V_s} = 0$$

$$I_c = h_{fe} I_b + h_{oe} V_c$$

$$\frac{I_c}{V_c} = h_{fe}\frac{I_b}{V_c} + h_{oe}$$

when $V_s = 0$, $\quad R_s . I_b + h_{ie} . I_b + h_{re} V_c = 0.$

$$\frac{I_b}{V_c} = -\frac{h_{re}}{R_s + h_{ie}}$$

$$\therefore Y_O = h_{oe} - \frac{h_{re}\, h_{fe}}{R_s + h_{ie}}$$

Voltage amplification taking into account source impedance (R_S) is given by

$$A_{VS} = \frac{V_c}{V_s} = \frac{V_c}{V_b} * \frac{V_b}{V_S} \qquad \left(V_b = \frac{V_s}{R_s + Z_i} * Z_i\right)$$

$$= A_V \cdot \frac{Z_i}{Z_i + R_s}$$

$$= \frac{A_i Z_L}{Z_i + R_s} \qquad \text{(E-3.68)}$$

A_v is the voltage gain for an ideal voltage source ($R_v = 0$).

Consider input source to be a current source I_S in parallel with a resistance R_S as shown in **fig.3.68**.

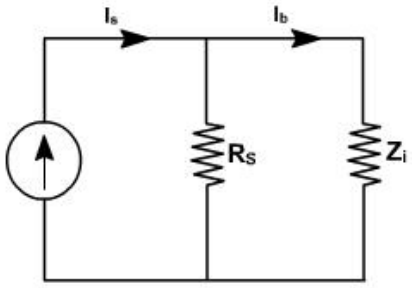


Fig. 3.68

In this case, overall current gain A_{IS} is defined as

$$A_{I_s} = \frac{I_L}{I_s}$$

$$= -\frac{I_c}{I_s}$$

$$= -\frac{I_c}{I_b} * \frac{I_b}{I_s} \qquad \left(I_b = \frac{I_s * R_s}{R_s + Z_i}\right)$$

$$= A_I * \frac{R_s}{R_s + Z_i}$$

$$\text{If } R_s \to \infty, \quad A_{Is} \to A_I$$

(E-3.69)

To analyze multistage amplifier the h-parameters of the transistor used are obtained from manufacture data sheet. The manufacture data sheet usually provides h-parameter in CE configuration. These parameters may be converted into CC and CB values. For example **fig.3.69** h_{rc} in terms of CE parameter can be obtained as follows.

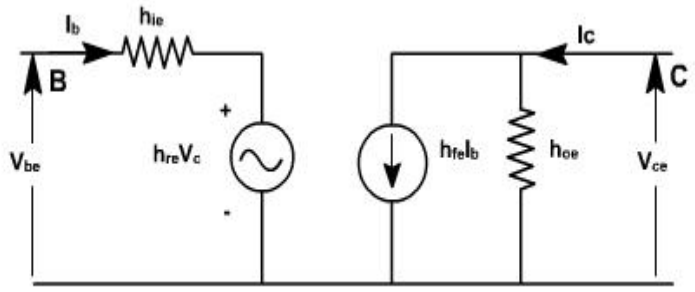


Fig.3.69

For CE transistor configuaration

$V_{be} = h_{ie} I_b + h_{re} V_{ce}$

$I_c = h_{fe} I_b + h_{oe} V_{ce}$ (E-3.70)

The circuit can be redrawn like CC transistor configuration as shown in **fig.3.70**.

$V_{bc} = h_{ie} I_b + h_{rc} V_{ec}$

$I_c = h_{fe} I_b + h_{oe} V_{ec}$ (E-3.71)

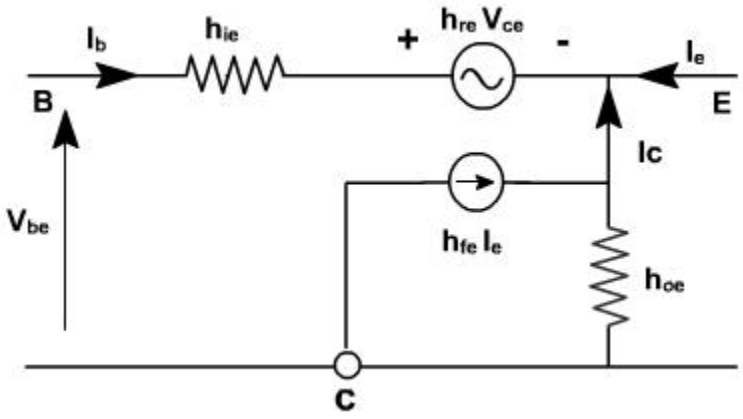


Fig.3.70

$$h_{rc} = \left.\frac{V_{be}}{V_{ec}}\right|_{I_b=0}$$

$$= \left.\frac{V_{be}+V_{ec}}{V_{ec}}\right|_{I_b=0}$$

$$= \left.\left(\frac{V_{be}}{V_{ec}}+1\right)\right|_{I_b=0}$$

Since $I_b = 0$, $V_{be} = h_{re} V_c = -h_{re} V_{ec}$

$$\therefore h_{rc} = 1+\left(\frac{h_{re} V_{ec}}{V_{ec}}\right)$$

$$= 1 - h_{re} \qquad \text{(E-3.72)}$$

Similarly

$$h_{fc} = \left.\frac{I_e}{I_b}\right|_{V_{ec}=0} = \left.\frac{-(I_b+I_C)}{I_b}\right|_{V_{ec}=0}$$

$$= -(1+h_{fe})$$

Example - 1

For the circuits shown in **fig. 3.71**. (CE–CC configuration) various h-parameters are given

h_{ie} = 2K, h_{fe} = 50, $h_{re} = 6 * 10^{-4}$, h_{oc}= 25 □ A/ V.

h_{ic} = 2K, h_{fe} = -51, h_{re} = 1, h_{oc} = 25 □ A/ V.

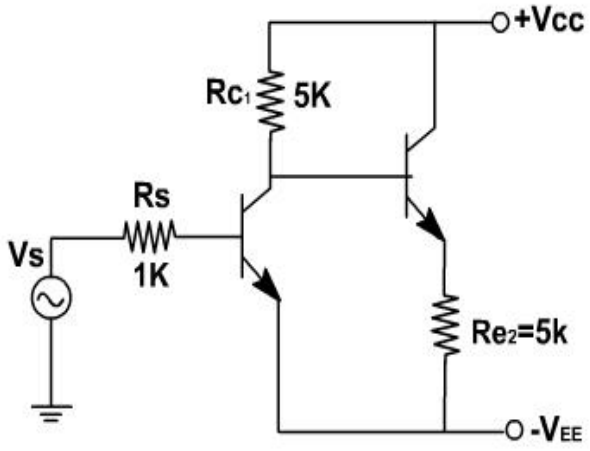


Fig.3.71

The small signal model of the transistor amplifier is shown in **fig. 3.72.**

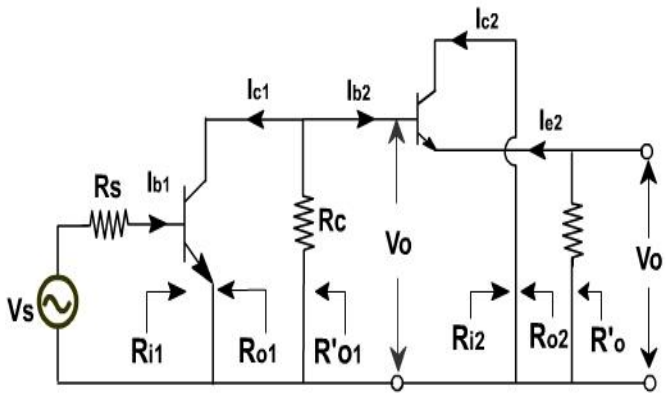


Fig.3.72

In the circuit, the collector resistance of first stage is shunted by the input impedence of last stage. Therefore the analysis is started with last stage. It is convenient; to first compute current gain, input impedence and voltage gain. Then output impedence is calculated starting from first stage and moving towards end.

$$A_{i2} = \frac{-h_{fe}}{1+h_{oc}Z_L} = \frac{51}{1+25*10^{-6}*5*10^3}$$
$$= 45.3$$

$$R_{i2} = h_{ic} + h_{rc}A_{i2}Z_L$$
$$= 2*10^3 + 1*45.3*5*10^3$$
$$= 228.5K \text{ (high input impedence)}$$

$$A_{V2} = \frac{V_0}{V_2} = \frac{A_i Z_L}{Z_{i2}}$$
$$= \frac{45.3*5}{228.5} = 0.99 \approx 1$$

$$R_{L1} = R_{C1} \| R_{i2}$$
$$= \frac{5*228.5}{5+228.5} = 4.9K$$

$$A_{i1} = -\frac{h_{fe}}{1+h_{oe}R_L} = \frac{-50}{1+25*10^{-6}*4.9*10^3}$$
$$= 44.5$$

$$R_{i1} = h_{ie} + h_{re}A_{i1}R_{L1}$$
$$= 2-6*10^{-4}*44.5*4.9$$
$$= 1.87K$$

Voltage gain of first stage is

$$A_{V1} = \frac{A_{i1}R_{L1}}{R_{i1}} = \frac{-44.5*4.9}{1.87}$$
$$= -116.6$$

$$Y_{o1} = h_{oe} - \frac{h_{fe}\,h_{re}}{h_{ie}+R_s}$$
$$= 25*10^{-6} - \frac{50*6*10^{-4}}{2*10^3+1*10^3}$$
$$= 15*10^{-6} \text{ mho}$$

$$R_{o1} = \frac{1}{Y_{o1}} = 66.7K$$

$$R'_{o1} = R_{o1} \| R_{c1}$$
$$= 66.7 \| 5$$
$$= 4.65K$$

The effective source resistance R'_{S2} for the second stage is $R_{01} \parallel R_{C1}$. Thus $R_{S2} = R'_{01} = 4.65K$

Frequency response

$$Y_{02} = h_{oe} - \frac{h_{fe}\, h_{rc}}{h_{ic} + R_{S2}}$$

$$= 25*10^{-6} - \frac{(-51)(1)}{2*10^{3} + 4.65*10^{3}}$$

$$= 7.70*10^{-3}\ \text{A/V}$$

$$R_{02} = \frac{1}{Y_{02}} = 130\Omega$$

$$R'_{02} = R_{02} \parallel R_{c2}$$

$$= 0.13 \parallel 5K$$

$$= 127\Omega$$

Overall current gain of the amplifier is A_i and is given by

$$A_i = -\frac{I_{e2}}{I_{b1}}$$

$$= -\frac{I_{e2}}{I_{b2}} * \frac{I_{b2}}{I_{c1}} * \frac{I_{c1}}{I_{b1}}$$

$$= -A_{i2} \cdot \frac{I_{b2}}{I_{c1}} * A_{i1}$$

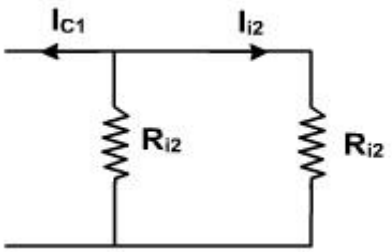


Fig. 3.73

The equivalent circuit of the amplifier is shown in **fig. 3.73**. From the circuit it is clear that the current i_{c1} is divided into two parts.

Therefore,

$$\frac{I_{b2}}{I_{c1}} = \frac{-R_{c1}}{R_{c1} + R_{i2}}$$

and

$$\therefore A_I = A_{I2} \cdot A_{I1} \cdot \frac{R_{c1}}{R_{c1}+R_{i2}}$$
$$= 45.3 * (-44.5) * \frac{5}{228.5+5} = -43.2$$
$$A_V = \frac{V_0}{V_I} = \frac{V_0}{V_2} \cdot \frac{V_2}{V_1}$$
$$= A_{V2} \cdot A_{V1}$$
$$= 0.99 * (-11.6)$$
$$= 115$$

Overall voltage gain of the amplifier is given by

$$A_{VS} = \frac{V_0}{V_S} = A_V \cdot \frac{R_{i1}}{R_{i1}+R_S}$$
$$= -115 * \frac{1.87}{1.87+1}$$
$$= -75.3$$

3.13 Embers-Moll Model

(Developed by Ebers and Moll in 1954) Also called the "Coupled Diode Model" Two diodes: The EB and CB diode

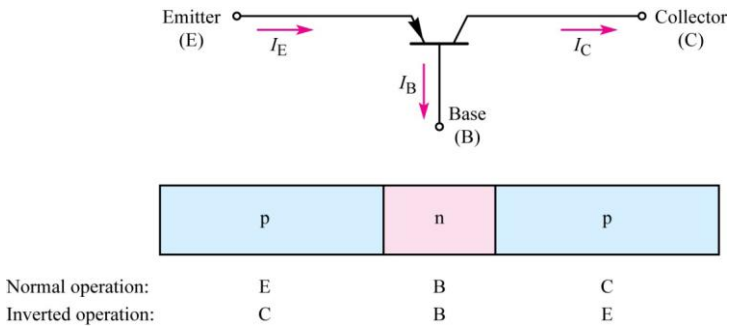


Fig 3.76

Charge distribution in the Base:

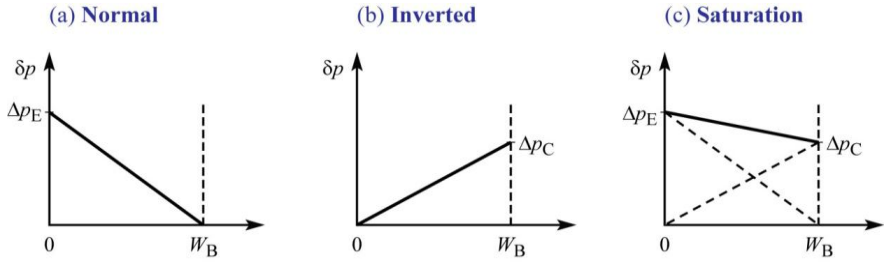


Fig 3.77

Under ***Normal*** operation (Subscript **N**), it is $\Delta p_C = 0$

Using $\Delta p_C = 0$ and the results of the last chapter, we have:

$$I_{EN} = a\,\Delta p_E \qquad I_{CN} = b\,\Delta p_E$$

Under ***Inverted*** operation (Subscript **I**), it is $\Delta p_E = 0$

Using $\Delta p_E = 0$ and the results of the last chapter, we have:

$$I_{EI} = -b\,\Delta p_C \qquad I_{CI} = -a\,\Delta p_C$$

In Eqs. (1) and (2), it is

$$a = e\,A\,\frac{D_p}{L_p}\coth\frac{W_B}{L_p}$$

$$b = e\,A\,\frac{D_p}{L_p}\operatorname{cosech}\frac{W_B}{L_p}$$

Superposition of the two currents:

$$\begin{aligned} I_E &= I_{EN} + I_{EI} \\ &= a\,\Delta p_E - b\,\Delta p_C \\ &= a\,p_{n0}\,(e^{eV_{EB}/kT} - 1) - b\,p_{n0}\,(e^{eV_{CB}/kT} - 1) \end{aligned}$$

$$\begin{aligned} I_C &= I_{CN} + I_{CI} \\ &= b\,\Delta p_E - a\,\Delta p_C \\ &= b\,p_{n0}\,(e^{eV_{EB}/kT} - 1) - a\,p_{n0}\,(e^{eV_{CB}/kT} - 1) \end{aligned}$$

The superposition of normal operation currents and inverted operation currents

yields two equations valid for any operating condition.

The four currents I_{EN}, I_{EI}, I_{CN}, and I_{CI} can also be written in the following w

Normal operation (V_{BE} = forward V_{CB} = reverse)

$$I_{EN} = I_{ES}\,(e^{eV_{EB}/kT} - 1)$$

$$I_{CN} = \alpha_N I_{EN} = \alpha_N I_{ES}\,(e^{eV_{EB}/kT} - 1)$$

α_N … current amplification in normal operation

Inverted operation (V_{EB} = reverse V_{CB} = forward)

$$I_{CI} = -I_{CS}\,(e^{eV_{CB}/kT} - 1)$$

$$I_{EI} = \alpha_I I_{CI} = -\alpha_I I_{CS}\,(e^{eV_{CB}/kT} - 1)$$

α_I … current amplification in inverted operation

Superposition of ***Normal*** and ***Inverted*** operation:
(Why are we allowed to superimpose? ➔ Linear processes)

$$I_E = I_{EN} + I_{EI} = I_{ES}\left(e^{eV_{EB}/kT} - 1\right) - \alpha_I I_{CS}\left(e^{eV_{CB}/kT} - 1\right)$$

$$I_C = I_{CN} + I_{CI} = \alpha_N I_{ES}\left(e^{eV_{EB}/kT} - 1\right) - I_{CS}\left(e^{eV_{CB}/kT} - 1\right)$$

We can construct an equivalent circuit from the Ebers-Moll Equations

Ebers-Moll equivalent circuit for **all** operating conditions:

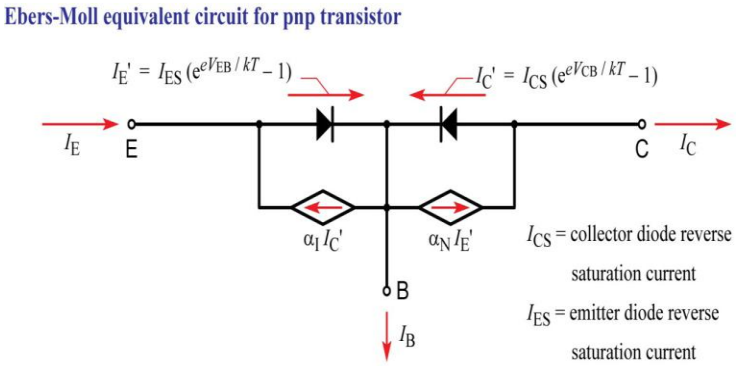


Fig 3.78

Ebers-Moll equivalent circuit for normal operating conditions:

CB diode is reverse biased → I 'C ≈ 0

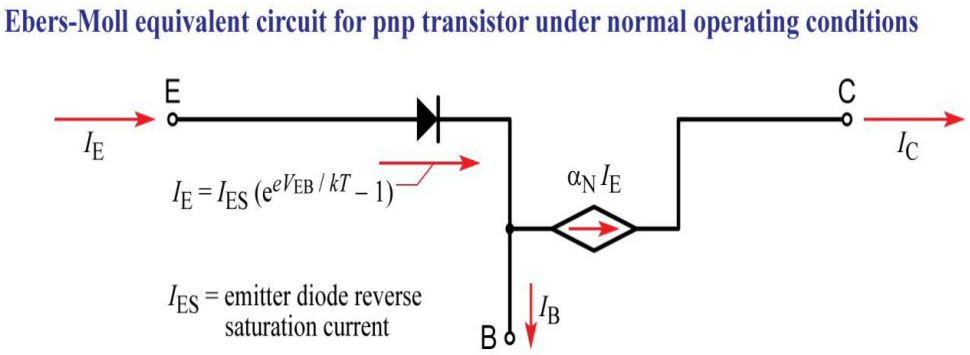


Fig 3.79

Ebers-Moll equivalent circuit is the bridge between internal device physics and electronic circuits.

Field Effect Transistor:

The field effect transistor is a semiconductor device, which depends for its operation on the control of current by an electric field. There are two of field effect transistors:

1. JFET (Junction Field Effect Transistor)
2. MOSFET (Metal Oxide Semiconductor Field Effect Transistor)

The FET has several advantages over conventional transistor.

1. In a conventional transistor, the operation depends upon the flow of majority and minority carriers. That is why it is called bipolar transistor. In FET the operation depends upon the flow of majority carriers only. It is called unipolar device.
2. The input to conventional transistor amplifier involves a forward biased PN junction with its inherently low dynamic impedance. The input to FET involves a reverse biased PN junction hence the high input impedance of the order of M-ohm.
3. It is less noisy than a bipolar transistor.
4. It exhibits no offset voltage at zero drain current.
5. It has thermal stability.
6. It is relatively immune to radiation.

The main disadvantage is its relatively small gain bandwidth product in comparison with conventional transistor.

4.2 Operation of FET:

Consider a sample bar of N-type semiconductor. This is called N-channel and it is electrically equivalent to a resistance as shown in **fig. 4.1**.

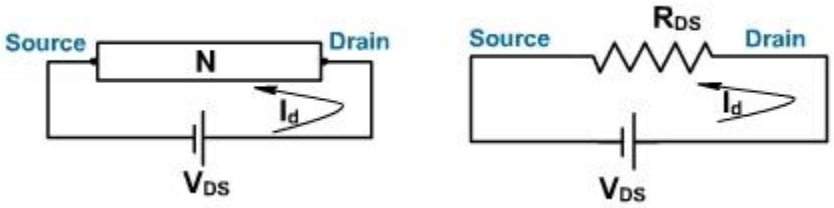


Fig. 4.1

Ohmic contacts are then added on each side of the channel to bring the external connection. Thus if a voltage is applied across the bar, the current flows through the channel.

The terminal from where the majority carriers (electrons) enter the channel is called source designated by S. The terminal through which majority carriers leaves the channel is called drain and designated by D. For an N-channel device, electrons are the majority carriers. Hence the circuit behaves like a dc voltage V_{DS} applied across a resistance R_{DS}. The resulting current is the drain current I_D. If V_{DS} increases, I_D increases proportionally.

Now on both sides of the n-type bar heavily doped regions of p-type impurity have been formed by any method for creating pn junction. These impurity regions are called gates (gate1 and gate2) as shown in **fig.4.**1.

Both the gates are internally connected and they are grounded yielding zero gate source voltage (V_{GS} =0). The word gate is used because the potential applied between gate and source controls the channel width and hence the current.

As with all PN junctions, a depletion region is formed on the two sides of the reverse biased PN junction. The current carriers have diffused across the junction, leaving only uncovered positive ions on the n side and negative ions on the p side. The depletion region width increases with the magnitude of reverse bias. The conductivity of this channel is normally zero because of the unavailability of current carriers.

The potential at any point along the channel depends on the distance of that point from the drain, points close to the drain are at a higher positive potential, relative to ground, then points close to the source. Both depletion regions are therefore subject to greater reverse voltage near the drain. Therefore the depletion region width increases as we move towards drain. The flow of electrons from source to drain is now restricted to the narrow channel between the no conducting depletion regions. The width of this channel determines the resistance between drain and source.

Consider now the behavior of drain current I_D vs drain source voltage V_{DS}. The gate source voltage is zero therefore $V_{GS}= 0$. Suppose that V_{DS} is gradually linearly increased linearly from 0V. I_D also increases.

Since the channel behaves as a semiconductor resistance, therefore it follows ohm's law. The region is called ohmic region, with increasing current, the ohmic voltage drop between the source and the channel region reverse biased the junction, the conducting portion of the channel begins to constrict and I_D begins to level off until a specific value of V_{DS} is reached, called the **pinch of voltage V_P**.

At this point further increase in V_{DS} do not produce corresponding increase in I_D. Instead, as V_{DS} increases, both depletion regions extend further into the channel, resulting in a no more cross section, and hence a higher channel resistance. Thus even though, there is more voltage, the resistance is also greater and the current remains relatively

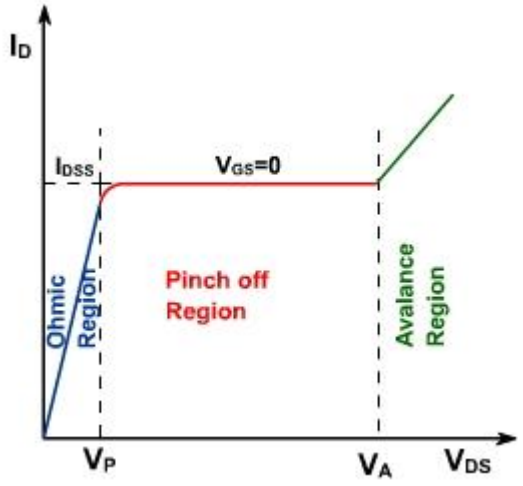


Fig. 4.2

constant. This is called pinch off or saturation region. The current in this region is maximum current that FET can produce and designated by I_{DSS}. (Drain to source current with gate shorted).

As with all pn junctions, when the reverse voltage exceeds a certain level, avalanche breakdown of pn junction occurs and I_D rises very rapidly as shown in **fig.4.2**.

Consider now an N-channel JFET with a reverse gate source voltage as shown in **fig.4.3**.

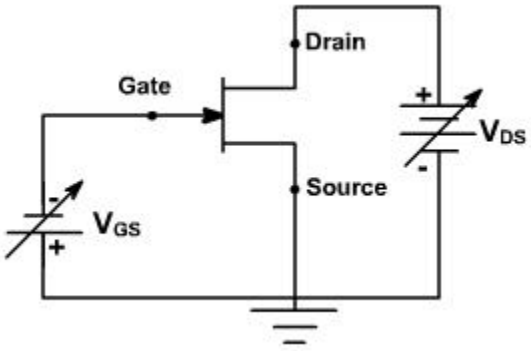


Fig. 4.3

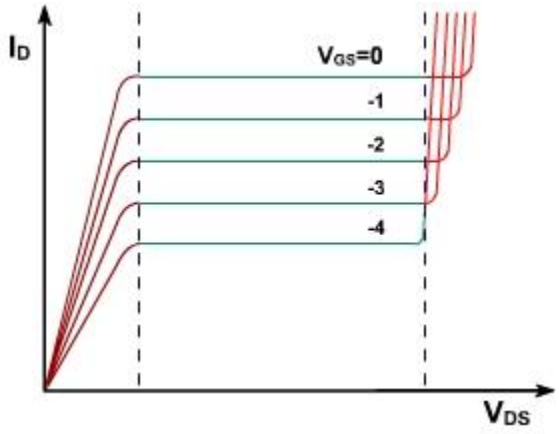


Fig. 4.5

The additional reverse bias, pinch off will occur for smaller values of $|V_{DS}|$, and the maximum drain current will be smaller. A family of curves for different values of V_{GS}(negative) is shown in **fig. 4.5**.

Suppose that V_{GS}= 0 and that due of V_{DS} at a specific point along the channel is +5V with respect to ground. Therefore reverse voltage across either p-n junction is now 5V. If V_{GS} is decreased from 0 to –1V the net reverse bias near the point is 5 - (-1) = 6V. Thus for any fixed value of V_{DS}, the channel width decreases as V_{GS} is made more negative.

Thus I_D value changes correspondingly. When the gate voltage is negative enough, the depletion layers touch each other and the conducting channel pinches off (disappears). In this case the drain current is cut off. The gate voltage that produces cut off is symbolized V_{GS}(off) . It is same as pinch off voltage.

Since the gate source junction is a reverse biased silicon diode, only a very small reverse current flows through it. Ideally gate current is zero. As a result, all the free electrons from the source go to the drain i.e. $I_D = I_S$. Because the gate draws almost negligible reverse current the input resistance is very high 10's or 100's of M ohm. Therefore where high input impedance is required, JFET is preferred over BJT. The disadvantage is less control over output current i.e. FET takes larger changes in input voltage to produce changes in output current. For this reason, JFET has less voltage gain than a bipolar amplifier.

4.3 Biasing the Field Effect Transistor

Transductance Curves: The transductance curve of a JFET is a graph of output current (I_D) vs input voltage (V_{GS}) as shown in **fig.4.6**.

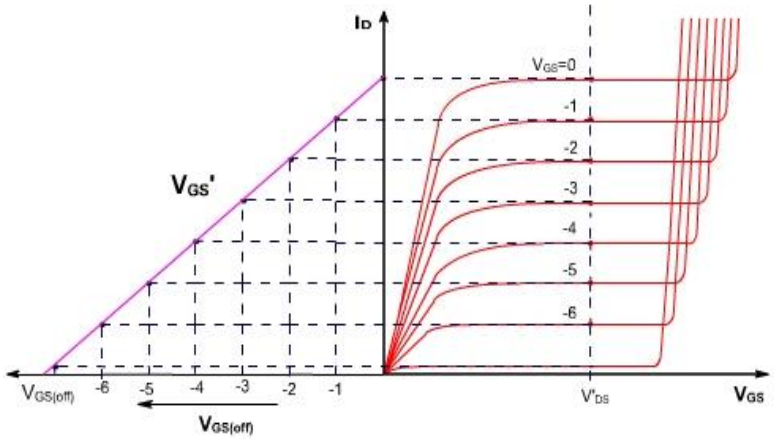


Fig. 4.6

By reading the value of I_D and V_{GS} for a particular value of V_{DS}, the transductance curve can be plotted. The transductance curve is a part of parabola. It has an equation of

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_{GS(off)}}\right)^2 \qquad \text{(E-4.1)}$$

Data sheet provides only I_{DSS} and V_{GS}(off) value. Using these values the transductance curve can be plotted.

Biasing the FET:

The FET can be biased as an amplifier. Consider the common source drain characteristic of a JFET. For linear amplification, Q point must be selected somewhere in the saturation region. Q point is selected on the basis of ac performance i.e. gain, frequency response, noise, power, current and voltage ratings.

4.3.1 Gate Bias:

Fig.4.7, shows a simple gate bias circuit.

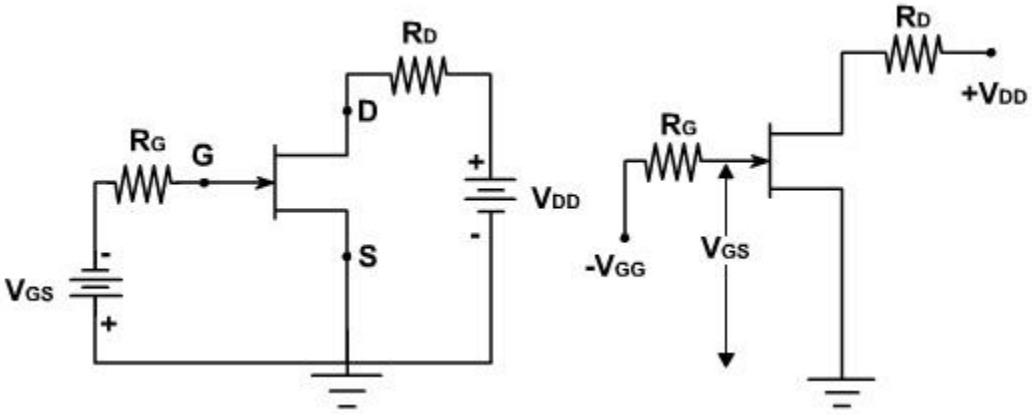


Fig. 4.7

Separate V_{GS} supply is used to set up Q point. This is the worst way to select Q point. The reason is that there is considerable variation between the maximum and minimum values of FET parameters e.g.

	I_{DSS}	V_{GS}(off)
Minimum	4mA	-2V

Maximum 13mA -8V

This implies that the minimum and maximum transductance curves are displaced as shown in **fig. 4.8**.

Gate bias applies a fixed voltage to the gate. This fixed voltage results in a Q point that is highly sensitive to the particular JFET used. For instance, if V_{GS}= -1V the Q point may very from Q_1 to Q_2 depending upon the JFET parameter is use.

At Q_1, I_D= 0.016 (1 - (1/8))^2 = 12.3 mA

At Q_2, I_D= 0.004 (1-(1/2))^2 = 1 mA.

The variation in drain current is very large.

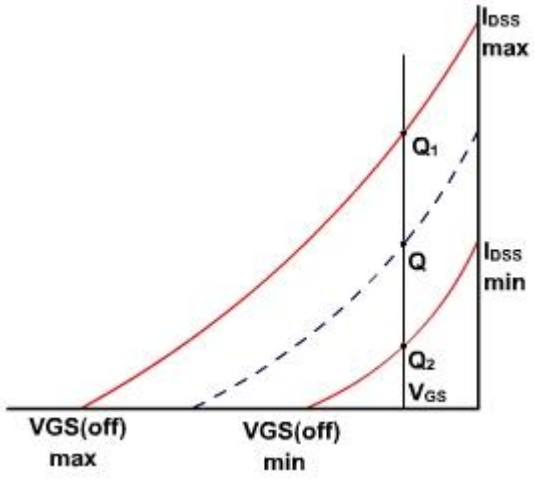


Fig. 4.8

4.3.2 Self Bias:

Fig.4.9, shows a self bias circuit another way to bias a FET. Only a drain supply is used and no gate supply. The idea is to use the voltage across R_S to produce the gate source reverse voltage.

This is a form of a local feedback similar to that used with bipolar transistors. If drain current increases, the voltage drop across R_S increases because the I_D R_S increases. This increases the gate source reverse voltage which makes the channel narrow and reduces the drain current. The overall effect is to partially offset the original increase in drain current. Similarly, if I_D decreases, drop across R_S decreases, hence reverse bias decreases and I_D increases.

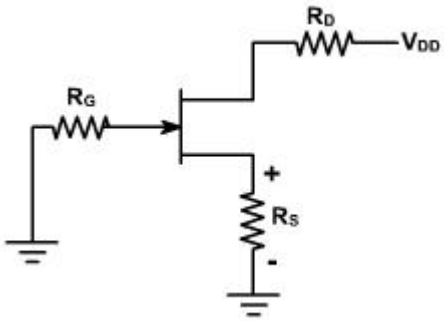


Fig. 4.9

Since the gate source junction is reverse biased, negligible gate current flows through R_G and so the gate voltage with respect to ground is zero.

$V_G = 0$;

The source to ground voltage equals the product of the drain current and the source resistance.

$\backslash V_S = I_D R_S$. (E-4.2)

The gate source voltage is the difference between the gate voltage and the source voltage.

$V_{GS} = V_G - V_S = 0 - I_D R_S$

$V_{GS} = -I_D R_S$. (E-4.3)

This means that the gate source voltage equals the negative of the voltage across the source resistor. The greater the drain current, the more negative the gate source voltage becomes.

Rearranging the equation:

$I_D = -V_{GS} / R_S$ (E-4.4)

The graph of this equation is called self base line a shown in **Fig.4.10**.

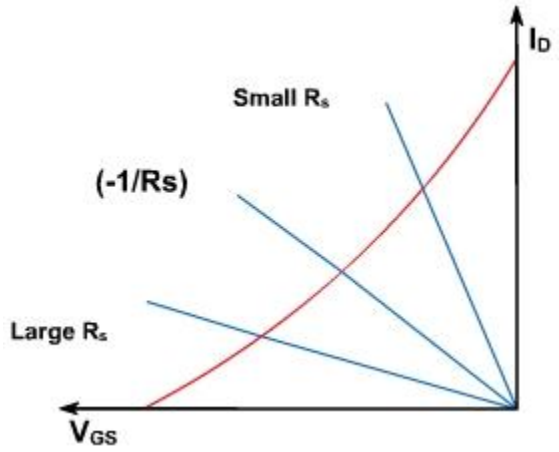


Fig:4.10

The operating point on transductance curve is the intersection of self bias line and transductance curve. The slope of the line is (-1 / R_S). If the source resistance is very large (-1 / R_S is small) then Q-point is far down the transductance curve and the drain current is small. When R_S is small, the Q point is far up the transductance curve and the drain current is large. In between there is an optimum value of R_S that sets up a Q point near the middle of the transductance curve.

The transductance curve varies widely for FET (because of variation in I_{DSS} and V_{GS}(off)) as shown in **fig. 4.11**. The actual curve may be in between there extremes. A and B are the optimum points for the two extreme curves. To find the optimum resistance R_S, so that Q-point is correct for all the curves, A and B points are joined such that it passes through origin.

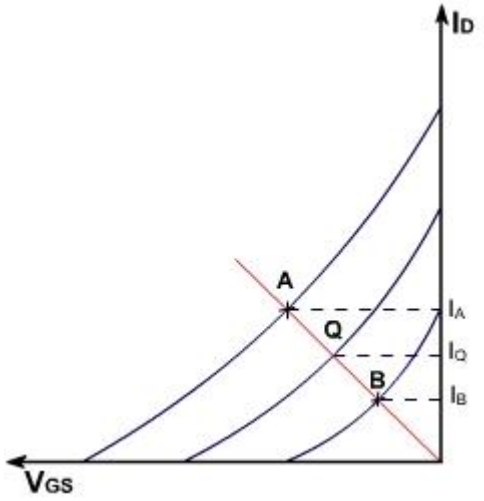


Fig. 4.11

The slope of this line gives the resistance value R_S(V_{GS} = -I_D R_S). The current I_Q is such that $I_A > I_Q > I_B$. Here A, Q and B all points are in straight line.

Consider the case where a line drawn to pass between points A and B does not pass through the origin. The equation V_{GS} = - I_D R_S is not valid. The equation of this line is $V_{GS} = V_{GG} – I_D R_S$.

Such a bias relationship may be obtained by adding a fixed bias to the gate in addition to the source self bias as shown in **fig.4.12**.

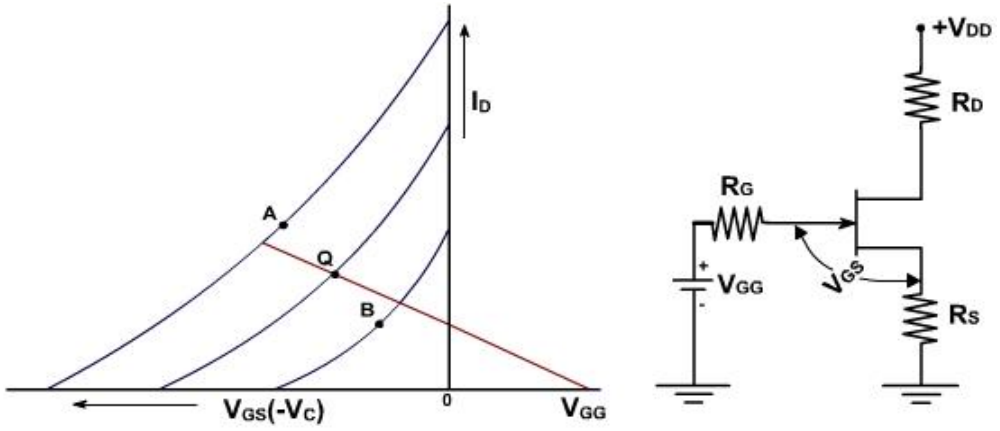


Fig.4.12

In this circuit.

$V_{GG} = R_S I_G + V_{GS} + I_D R_S$ (E-4.5)

Since $R_S I_G = 0$;

$V_{GG} = V_{GS} + I_D R_S$

or $V_{GS} = V_{GG} - I_D R_S$ (E-4.6)

4.3.3 Voltage Divider Bias :

The biasing circuit based on single power supply is shown in **fig. 4.13**. This is similar to the voltage divider bias used with a bipolar transistor.

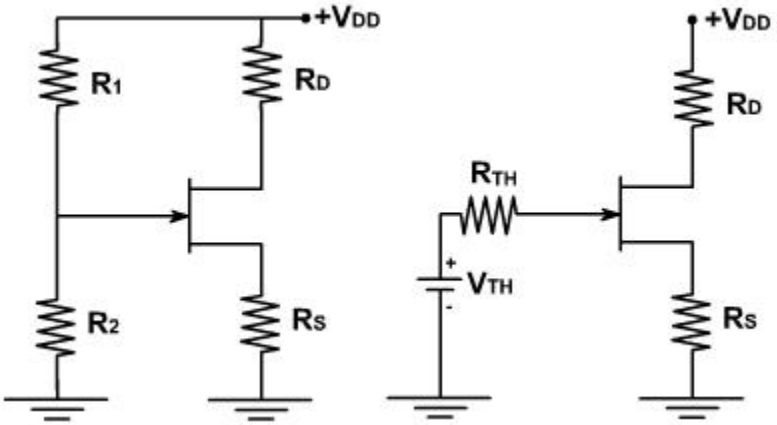


Fig. 4.13

The Thevenin voltage V_{TH} applied to the gate is

$$V_{TH} = \frac{R_2}{R_1 + R_2} V_{DD} \qquad \text{(E-4.7)}$$

The Thevenin resistance is given as

$$R_{TH} = \frac{R_2 R_1}{R_1 + R_2} \qquad \text{(E-4.8)}$$

The gate current is assumed to be negligible. V_{TH} is the dc voltage from gate to ground.

$$V_{TH} = V_{GS} + V_S \text{ (neglecting } I_G\text{)}$$
$$\therefore V_S = V_{TH} = V_{GS} \qquad \text{(E-4.9)}$$

The drain current ID is given by

$$I_D = \frac{V_{TH} - V_{GS}}{R_S} \qquad \text{(E-4.10)}$$

and the dc voltage from the drain to ground is $V_D = V_{DD} – I_D R_D$.

If V_{TH} is large enough to swamp out V_{GS} the drain current is approximately constant for any JFET as shown in **fig.4.14**.

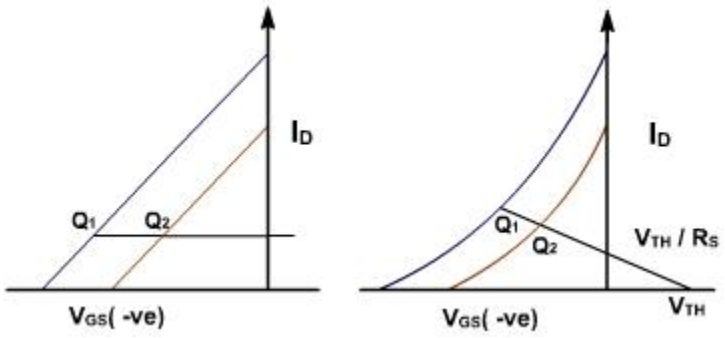


Fig. 4.14

There is a problem in JFET. In a BJT, V_{BE} is approximately 0.7V, with only minor variations from one transistor to other. In a FET, V_{GS} can vary several volts from one JFET to another. It is therefore, difficult to make V_{TH} large enough to swamp out V_{GS}. For this reason, voltage divider bias is less effective with, FET than BJT. Therefore, V_{GS} is not negligible. The current increases slightly from Q2 to Q1. However, voltage divider bias maintains I_D nearly constant.

Consider a voltage divider bias circuit shown in **fig.4.15**.

$V_{GS(min)} = -1, \quad V_{GS(max)} = -5V$

$V_{TH} = 15V$

$$I_{D(min)} = \frac{15-(-1)}{7.5K} = 2.13 \text{ mA}$$

$$I_{D(max)} = \frac{15-(-5)}{7.5K} = 2.67 \text{ mA}$$

Difference in $I_{D\,(min)}$ and $I_{D\,(max)}$ is less

$V_{D\,(max)} = 30 - 2.13 * 4.7 = 20$ V

$V_{D\,(min)} = 30 - 2.67 * 4.7 = 17.5$ V

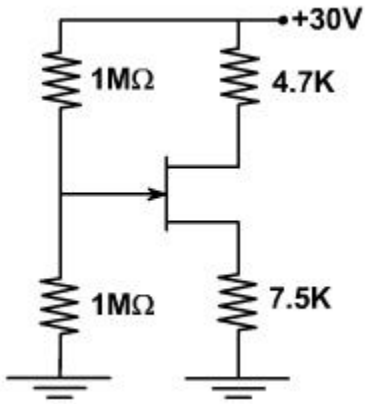


Fig.4.15

4.3.4 Current Source Bias:

This is another way to produce solid Q point. The aim is to produce a drain current that is independent of V_{GS}. Voltage divider bias and self bias attempt to do this by swamping out of variations in V_{GS}.

Using two power supplies:

The current source bias can be used to make I_D constant **fig.4.16**.

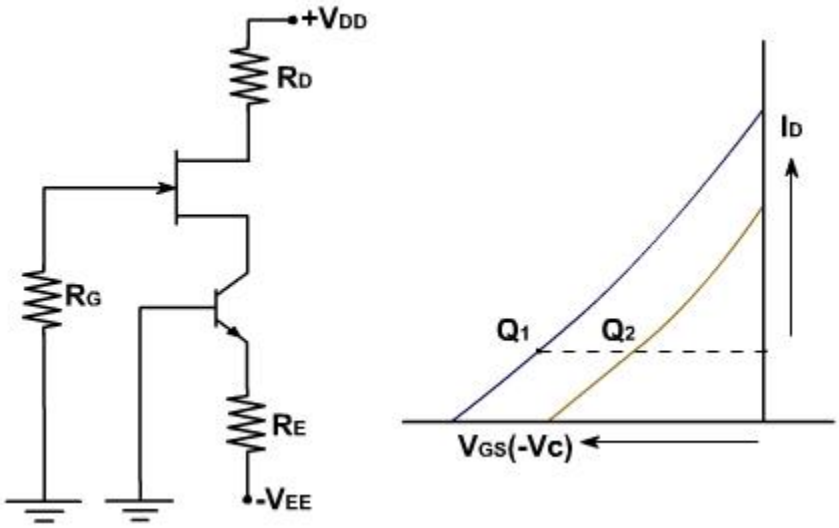


Fig. 4.16

The bipolar transistor is emitter biased; its collector current is given by

$I_C = (V_{EE} – V_{BE}) / R_E.$ (E-4.11)

Because the bipolar transistor acts like a current source, it forces the drain current to equal the bipolar collector current.

$I_D = I_C$

Since I_C is constant, both Q points have the same value of drain current. The current source effectively wipes out the influence of V_{GS}. Although V_{GS} is different for each Q point, it no longer influences the value of drain current.

Using One power supply:

When only a positive supply is available, the circuit shown in **fig. 5**, can be used to set up a constant drain current.

In this case, the bipolar transistor is voltage divider biased. Assuming a stiff voltage divider, the emitter and collector currents are constant for all bipolar transistors. This forces the FET drain current equal the bipolar collector current.

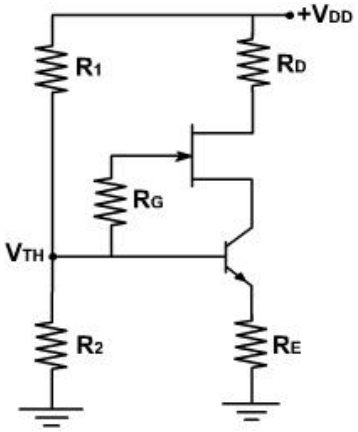


Fig. 4.17

$$V_{TH} = \frac{R_2 \, V_{DD}}{R_1 + R_2}$$

$$I_E = \frac{V_{TH} - V_{BE}}{R_E}$$

Since V_{TH} is constant, I_E is also constant

$$I_C = I_S = I_D = \text{constant}$$

4.4 Transductance:

The transductance of a FET is defined as

$$g_m = \left.\frac{\Delta I_D}{\Delta V_{gs}}\right|_{V_{DS}=0} \mu A/Volts \qquad \text{(E-4.12)}$$

Because the changes in I_D and V_{GS} are equivalent to ac current and voltage. This equation can be written as

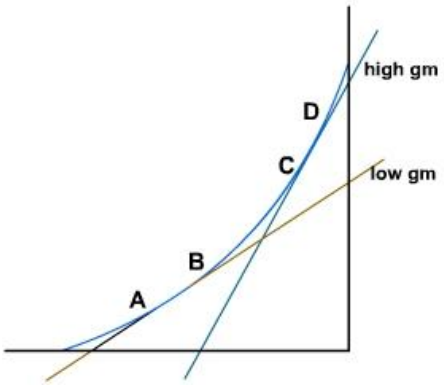


Fig. 4.18

$$g_m = \left.\frac{i_d}{V_{gs}}\right|_{V_{ds}=0} \qquad \text{(E-4.13)}$$

The unit of g_m is mho or siemems.

Typical value of g_m is 2000 m A / V.

The value of g_m can be obtained from the transductance curve as shown in **fig.4.18**.

If A and B points are considered, than a change in V_{GS} produces a change in I_D. The ratio of I_D and V_{GS} is the value of g_m between A and B points. If C and D points are considered, then same change in V_{GS} produces more change in I_D. Therefore, g_m value is higher. In a nutshell, g_m tells us how much control gate voltage has over drain current. Higher the value of g_m, the more effective is gate voltage in controlling gate current. The second parameter r_d is the drain resistance.

$$r_d = \left.\frac{v_{ds}}{i_d}\right|_{V_{gs}=0} \quad (r_d \text{ is negligible}) \qquad \text{(E-4.14)}$$

4.5 FET a amplifier

Similar to Bipolar junction transistor. JFET can also be used as an amplifier. The ac equivalent circuit of a JFET is shown in **fig.4.19**.

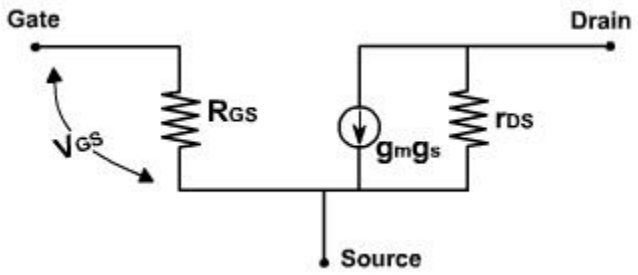


Fig. 4.19

The resistance between the gate and the source R_{GS} is very high. The drain of a JFET acts like a current source with a value of $g_m V_{gs}$. This model is applicable at low frequencies.

From the ac equivalent model

$$i_d = g_m V_{gs} + \frac{V_{ds}}{r_d}$$

$$\text{When } i_d = 0, \quad \frac{V_{ds}}{V_{gs}} = -g_m r_d \qquad \text{(E-4.15)}$$

The amplification factor μ for FET is defined as

$$\mu = \left.\frac{v_{ds}}{v_{gs}}\right|_{i_d=0} \quad \therefore \mu = g_m r_d \qquad \text{(E-4.16)}$$

When $V_{GS} = 0$, g_m has its maximum value. The maximum value is designated as $g_{mo.}$

Again consider the equation,

$$I_D = I_{DSS}\left[1-\frac{V_{GS}}{V_{GS(off)}}\right]^2$$

$$g_m = \frac{\partial I_D}{\partial V_{GS}} = 2I_{DSS}\left[1-\frac{V_{GS}}{V_{GS(off)}}\right]\left[\frac{-1}{V_{GS(off)}}\right]$$

$$g_m = \frac{-2I_{DSS}}{V_{GS(Off)}}\left[1-\frac{V_{GS}}{V_{GS(off)}}\right] \qquad \text{(E-4.17)}$$

When $V_{GS}=0$, $g_m = g_{mo} = \frac{-2I_{DSS}}{V_{GS(off)}}$

$$\therefore g_m = g_{mo}\left[1-\frac{V_{GS}}{V_{GS(off)}}\right] \qquad \text{(E-4.18)}$$

As V_{GS} increases, gm decreases linearly.

$$V_{GS(off)} = \frac{-2I_{DSS}}{g_{mo}} \qquad \text{(E-4.19)}$$

Measuring I_{DSS} and g_m, $V_{GS(off)}$ can be determined

FET as Amplifier:

Fig.4.20, shows a common source amplifier.

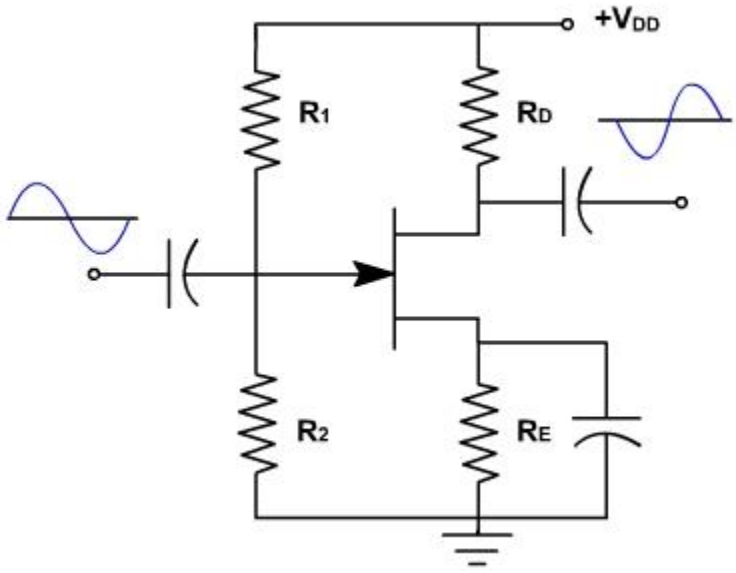


Fig. 4.20

When a small ac signal is coupled into the gate it produces variations in gate source voltage. This produces a sinusoidal drain current. Since an ac current flows through the drain resistor. An amplified ac voltage is obtained at the output. An increase in gate source voltage produces more drain current, which means that the drain voltage is decreasing. Since the positive half cycle of input voltage produces the negative half cycle of output voltage, we get phase inversion in a CS amplifier.

The ac equivalent circuit is shown in **fig.4.21**.

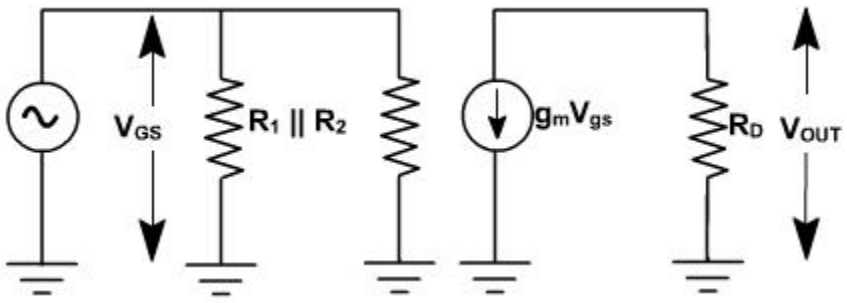


Fig. 4.22

The ac output voltage is

$v_{out} = - g_m v_{gS} R_D$ (E-4.20)

Negative sign means phase inversion. Because the ac source is directly connected between the gate source terminals therefore ac input voltage equals

$V_{in} = V_{gs}$ (E-4.21)

The voltage gain is given by

$$A_V = \frac{v_{out}}{v_{in}} = - g_m R_D$$

A_V = unloaded voltage gain (E-4.22)

The further simplified model of the amplifieris shown in **fig.4.23**.

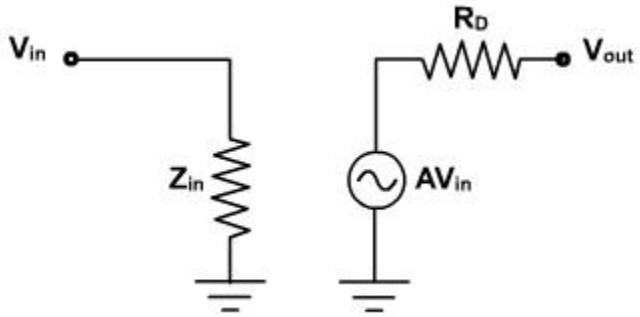


Fig. 4.23

Z_{in} is the input impedance. At low frequencies, this is parallel combination of $R_1 || R_2 || R_{GS}$. Since R_{GS} is very large, it is parallel combination of R_1 & R_2. A V_{in} is output voltage and R_D is the output impedance.

Because of nonlinear transductance curve, a JFET distorts large signals, as shown in **fig.4.24**.

Given a sinusoidal input voltage, we get a non-sinusoidal output current in which positive half cycle is elongated and negative cycle is compressed. This type of distortion is called Square law distortion because the transductance curve is parabolic.

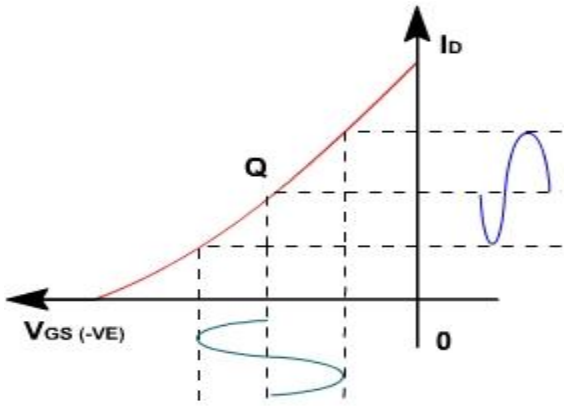


Fig.4.24

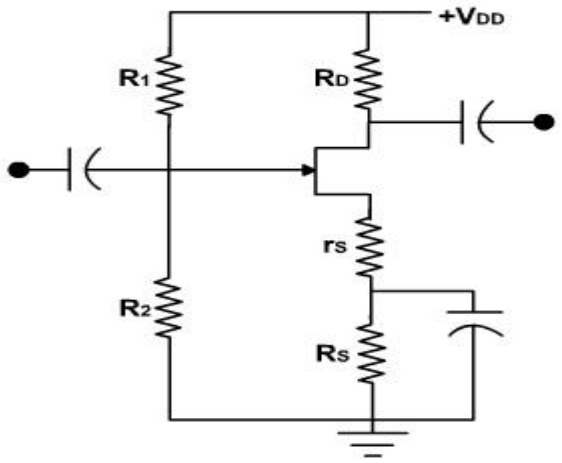


Fig. 2.25

This distortion is undesirable for an amplifier. One way to minimize this is to keep the signal small. In that case a part of the curve is used and operation is approximately linear. Some times swamping resistor is used to minimize distortion and gain constant. Now the source is no longer ac ground as shown in **fig. 4.25**.

The drain current through r_S produces an ac voltage between the source and ground. If r_S is large enough the local feedback can swamp out the non-linearity of the curve. Then the voltage gain approaches an ideal value of R_D / r_S.

Since R_{GS} approaches infinity therefore, all the drain current flows through r_S producing a voltage drop of $g_m V_{gS} r_S$. The ac equivalent circuit is shown in **fig. 4.26**.

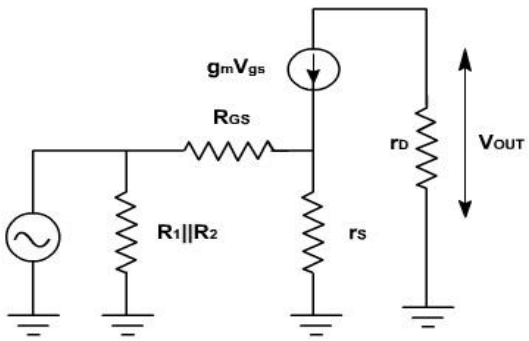


Fig. 4.26

$$v_{gs} + g_m\ v_{gs}.\ r_s - v_{in} = 0$$

$$v_{in} = (1 + g_m r_s)\ v_{gs}$$

$$v_{out} = - g_m R_D\ v_{gs}$$

$$A = \frac{- g_m\ R_D}{1 + g_m\ r_s} = \frac{-R_D}{r_s + 1/g_m}$$

(E-4.23)

The voltage gain reduces but voltage gain is less effective by change in g_m. r_S must be greater than $1 / g_m$ only then

$$v_{gs} = - \frac{R_D}{r_S}$$

(E-4.24)

4.6 JFET Applications

Example-1:

Determine gm for an n-channel JFET with characteristic curve shown in **fig. 4.27**

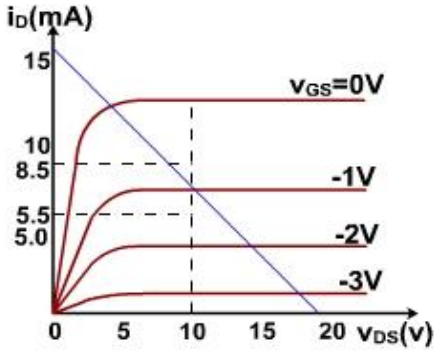


Fig. 4.27

Solution:

We select an operating region which is approximately in the middle of the curves; that is, between v_{GS} = -0.8 V and v_{GS} = -1.2 V; i_D = 8.5mA and i_D = 5.5 mA. Therefore, the transductance of the JFET is given by

$$g_m = \left.\frac{\Delta i_D}{\Delta v_{GS}}\right|_{v_{GS}=\text{constant}} = 7.5\ m\Omega^{-1} \qquad \text{(E-4.25)}$$

Design of JFET amplifier:

To design a JFET amplifier, the Q point for the dc bias current can be determined graphically. The dc bias current at the Q point should lie between 30% and 70% of I_{DSS}. This locates the Q point in the linear region of the characteristic curves.

The relationship between i_D and v_{GS} can be plotted on a dimensionless graph (i.e., a normalized curve) as shown in **fig.4.28** .

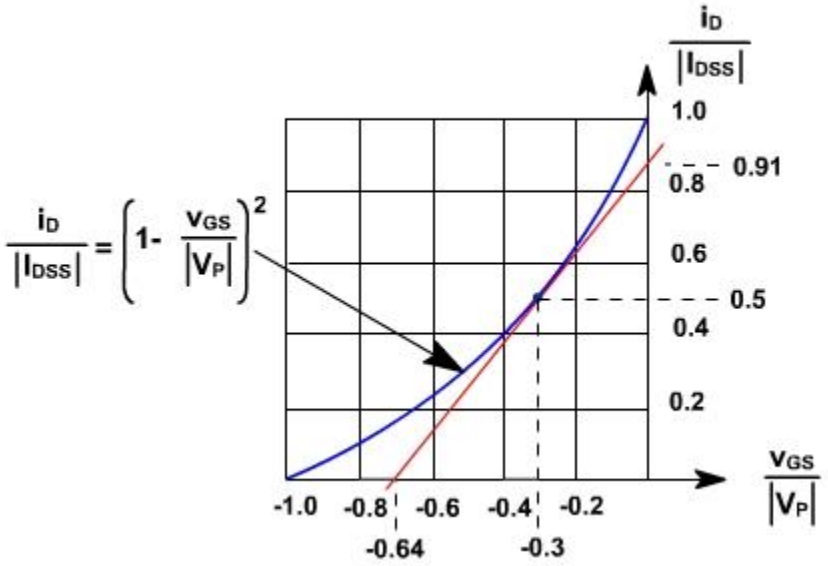


Fig. 4.28

The vertical axis of this graph is i_D / I_{DSS} and the horizontal axis is v_{GS} / V_P. The slope of the curve is g_m.

A reasonable procedure for locating the quiescent point near the center of the linear operating region is to select $I_{DQ} \approx I_{DSS} / 2$ and $V_{GSQ} \approx 0.3V_P$. Note that this is near the midpoint of the curve. Next we select $v_{DS} \approx V_{DD} / 2$. This gives a wide range of values for v_{ds} that keep the transistor in the pinch –off mode.

The transductance at the Q-point can be found from the slope of the curve of **fig.4.28** and is given by

$$g_m = \frac{1.41\ I_{DSS}}{V_P}$$

Example-2

Determine g m for a JFET where I_{DSS} = 7 mA, V_P = -3.5 V and V_{DD} = 15V. Choose a reasonable location for the Q-point.

Solution:

Let us select the Q-point as given below:

$$I_{DQ} = \frac{I_{DSS}}{2} = 3.5 \text{ mA}$$
$$V_{DSQ} = \frac{V_{DD}}{2} = 7.5 \text{ V}$$
$$V_{GSQ} = 0.3V_P = -1.05 \text{V}$$

The transconductance, g_m, is found from the slope of the curve at the point $i_D / I_{DSS} = 0.5$ and $v_{GS} / V_P = 0.3$. Hence,

$$g_m = \frac{1.41\ I_{DSS}}{V_P} = 2840\ \mu\Omega^{-1}$$

4.6.1 JEFT as Analog Switch:

JFET can be used as an analog switch as shown in **fig.4.29**. It is the major application of a JFET. The idea is to use two points on the load line: cut off and saturation. When JFET is cut off, it is like an open switch. When it is saturated, it is like a closed switch.

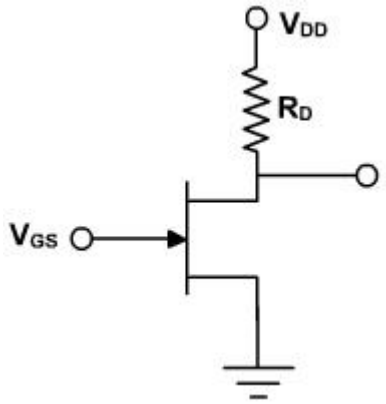


Fig. 4.29

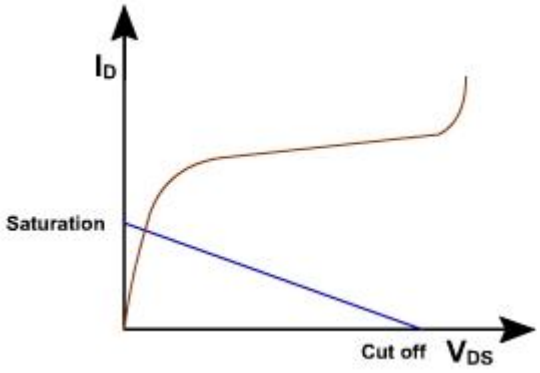


Fig. 4.30

When $V_{GS} = 0$, the JFET is saturated and operates at the upper end of the load line. When V_{GS} is equal to or more negative than $V_{GS}(off)$, it is cut off and operates at lower end of the load line (open and closed switch).This is shown in **fig.** 4.30.

Only these two points are used for operation when used as a switch. The JFET is normally saturated well below the knee of the drain curve. For this reason the drain current is much smaller than I_{DSS} .

4.6.2 FET as a Shunt Switch:

FET can be used as a shunt switch as shown in **fig.4.31**. When V_{cont}=0, the JFT is saturated and the switch is closed When V_{cont} is more negative FET is like an open switch. The equivalent circuit is also shown in

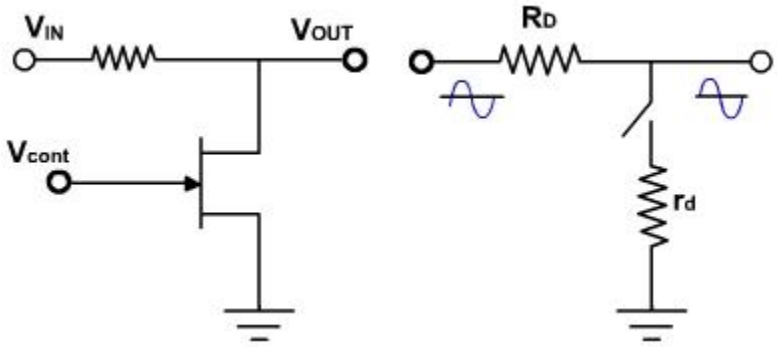


Fig.4.31

4.6. 3 FET as a series switch:

JFET can also be used as series switch as shown in **fig. 4.32**. When control is zero, the FET is a closed switch. When V_{con}= negative, the FET is an open switch. It is better than shunt switch.

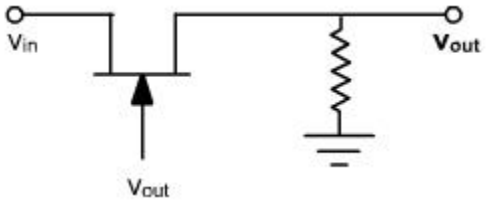


Fig. 4.32

4.7 Insulated-gate field-effect transistors (MOSFET)

The insulated-gate field-effect transistor (IGFET), also known as the metal oxide field effect transistor (MOSFET), is a derivative of the field effect transistor (FET). Today, most transistors are of the MOSFET type as components of digital integrated circuits. Though discrete BJT's are more numerous than discrete MOSFET's. The MOSFET transistor count within an integrated circuit may approach hundreds of a million. The dimensions of individual MOSFET devices are under a micron, decreasing every 18 months. Much larger MOSFET's are capable of switching nearly 100 amperes of current at low voltages; some handle nearly 1000 V at lower currents. These devices occupy a good fraction of a square centimeter of silicon. MOSFET's find much wider application than JFET's. However, MOSFET power devices are not as widely used as bipolar junction transistors at this time. The MOSFET has source, gate, and drain terminals like the FET. However, the gate lead does not make a direct connection to the silicon compared with the case for the FET. The MOSFET gate is a metallic or polysilicon layer atop a silicon dioxide insulator. The gate bears a resemblance to a metal oxide semiconductor (MOS) capacitor in Figure . When charged B the plates of the capacitor take on the charge polarity of the respective battery terminals. The lower plate is P-type silicon from which electrons are repelled by the negative (-) battery terminal toward the oxide, and attracted by the positive (+) top plate. This excess of electrons near the oxide creates an inverted (excess of electrons) channel under the oxide. This channel is also accompanied by a depletion region isolating the channel from the bulk silicon substrate. In Figure (a) the MOS capacitor is placed between a pair of N-type diffusions in a Ptypesubstrate. With no charge on the capacitor, no bias on the gate, the N-type diffusions, the source and drain, remain electrically isolated. A positive bias applied to the gate, charges the capacitor (the gate). The gate atop the oxide takes on a positive charge from the gate bias battery. The P-type substrate below the gate takes on a negative charge. An inversion region with an excess of electrons forms below the gate

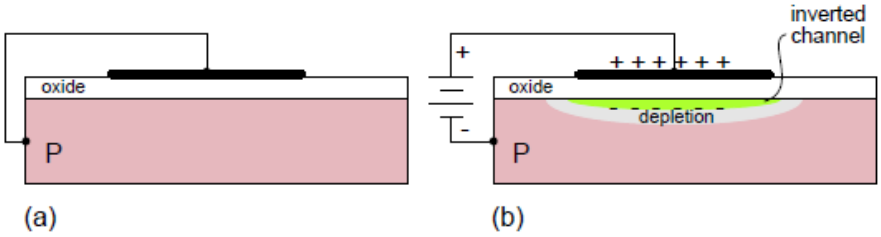
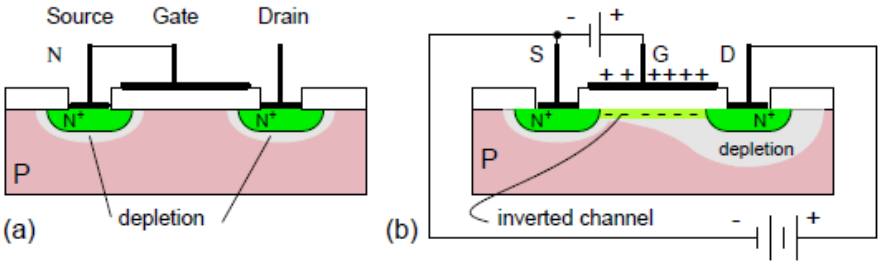


Figure 2.45: *N-channel MOS capacitor: (a) no charge, (b) charged.*



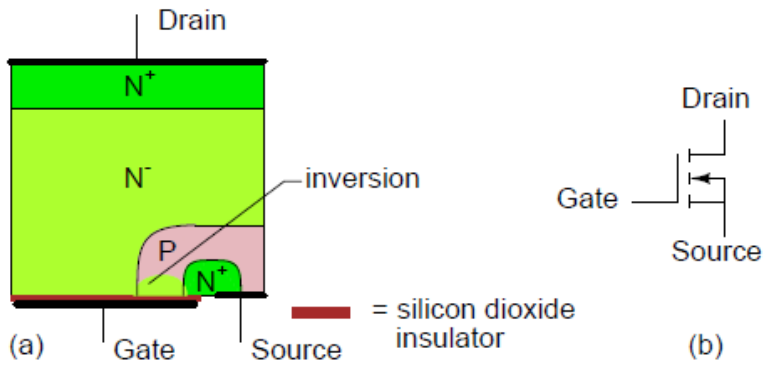
4.31 N-channel MOSFET (enhancement type): (a) 0 V gate bias, (b) positive gate bias.

oxide. This region now connects the source and drain N-type regions, forming a continuous N-region from source to drain. Thus, the MOSFET, like the FET is a unipolar device. One type of charge carrier is responsible for conduction. This example is an N-channel MOSFET. Conduction of a large current from source to drain is possible with a voltage applied between these connections. A practical circuit would have a load in series with the drain battery in Figure

The MOSFET described above in Figure is known as an enhancement mode MOSFET. The non-conducting, off, channel is turned on by enhancing the channel below the gate by application of a bias. This is the most common kind of device. The other kind of MOSFET will not be described here. See the Insulated-gate field-effect transistor chapter for the depletion mode device. The MOSFET, like the FET, is a voltage controlled device. A voltage input to the gate controls the flow of current from source to drain. The gate does not draw a continuous current. Though, the gate draws a surge of current to charge the gate capacitance. The cross-section of an N-channel discrete MOSFET is shown in Figure (a). Discrete devices are usually optimized for high power switching. The N+ indicates that the source and drain are heavily N-type doped. This minimizes resistive losses in the high current path from source to drain. The N− indicates light doping. The P-region under the gate, between source and drain can be inverted by application of a positive bias voltage. The doping

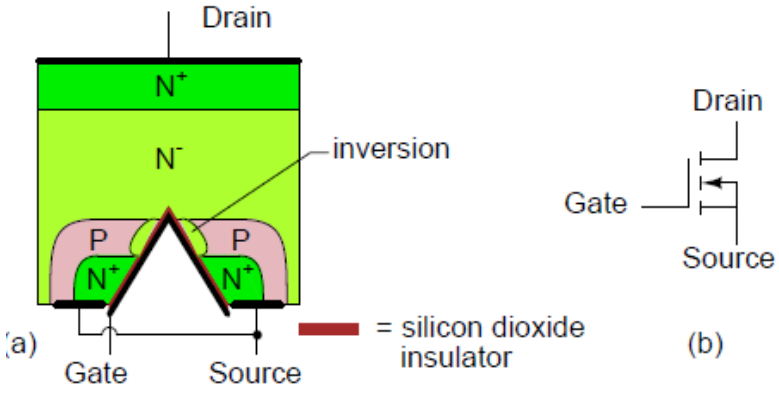
profile is a cross-section, which may be laid out in a serpentine pattern on the silicon die. This greatly increases the area, and consequently, the current handling ability.

The MOSFET schematic symbol in Figure (b) shows a "floating" gate, indicating no



4.32 N-channel MOSFET (enhancement type): (a) Cross-section, (b) schematic symbol

direct connection to the silicon substrate. The broken line from source to drain indicates that this device is off, not conducting, with zero bias on the gate. A normally "off " MOSFET is an enhancement mode device. The channel must be enhanced by application of a bias to the gate for conduction. The "pointing" end of the substrate arrow corresponds to P-type material, which points toward an N-type channel, the "non-pointing" end. This is the symbol for an N-channel MOSFET. The arrow points in the opposite direction for a P-channel device (not shown). MOSFET's are four terminal devices: source, gate, drain, and substrate. The substrate is connected to the source in discrete MOSFET's, making the packaged part a three terminal device. MOSFET's, that are part of an integrated circuit, have the substrate common to all devices, unless purposely isolated. This common connection may be bonded out of the die for connection to a ground or power supply bias voltage.



4.33 N-channel "V-MOS" transistor: (a) Cross-section, (b) schematic symbol

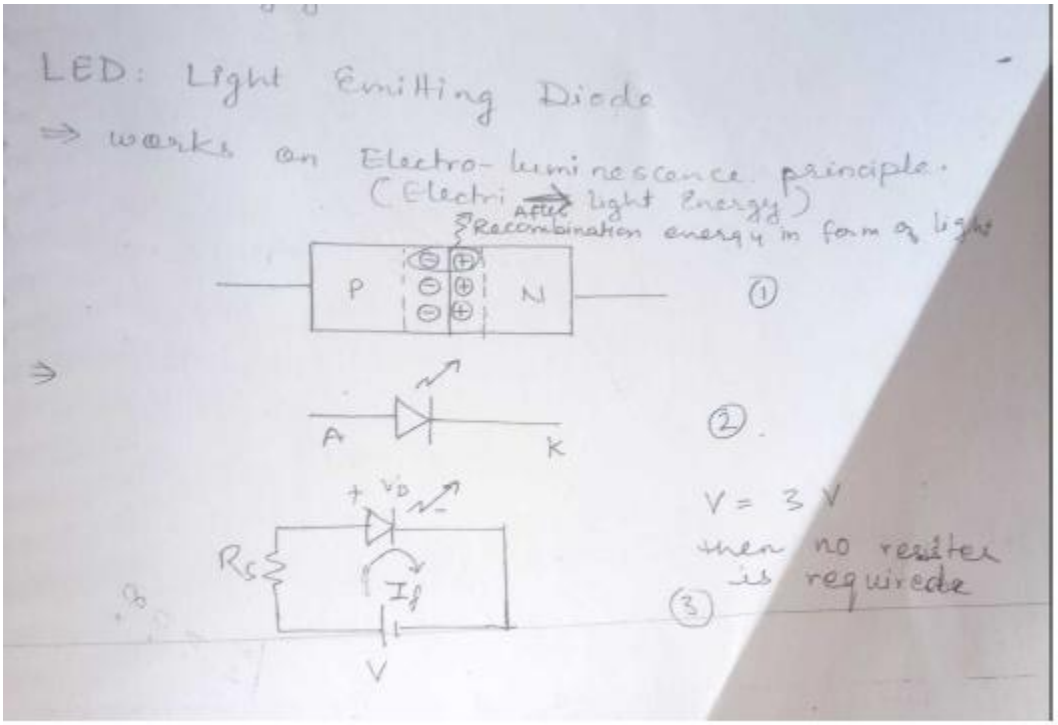
The *V-MOS* device in (Figure) is an improved power MOSFET with the doping profile arranged for lower on-state source to drain resistance. VMOS takes its name from the V-shaped gate region, which increases the cross-sectional area of the source-drain path. This minimizes losses and allows switching of higher levels of power. UMOS, a variation using a U-shaped grove, is more reproducible in manufacture.

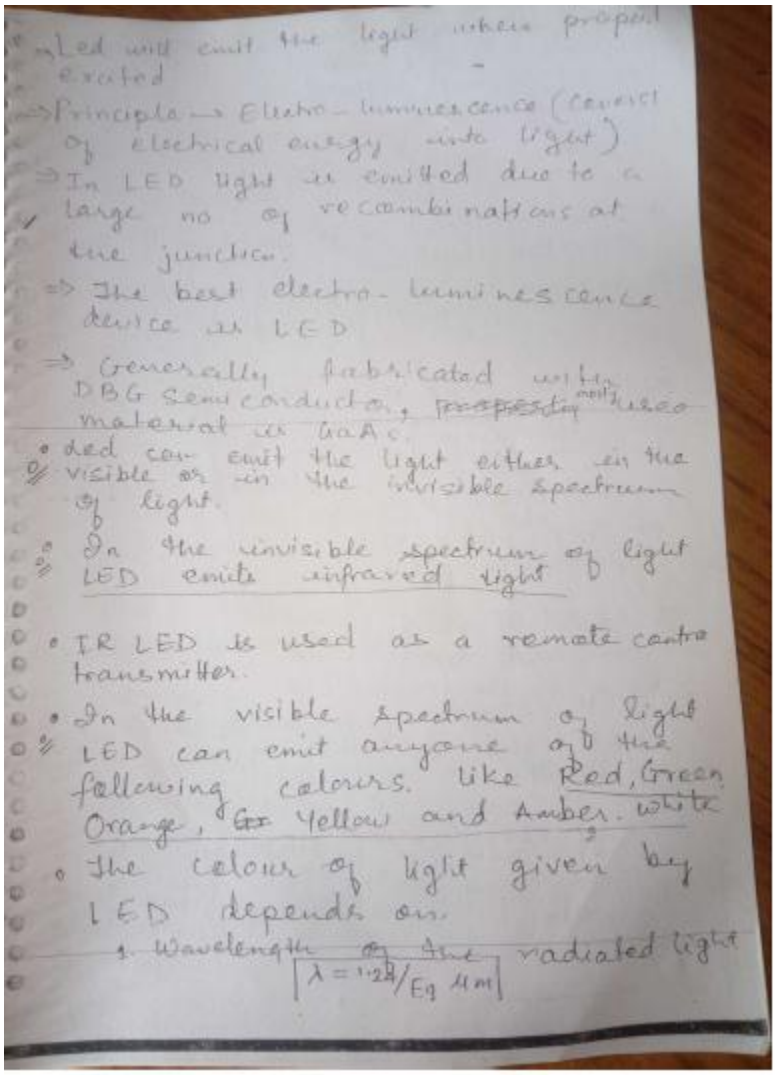
• MOSFET's are unipoar conduction devices, conduction with one type of charge carrier, like a FET, but unlike a BJT.

• A MOSFET is a voltage controlled device like a FET. A gate voltage input controls the source to drain current.

• The MOSFET gate draws no continuous current, except leakage. However, a considerable initial surge of current is required to charge the gate capacitance.

Light Emitting Diode





LED will emit the light when proper excited

=> Principle → Electro-luminescence (conversion of electrical energy into light)

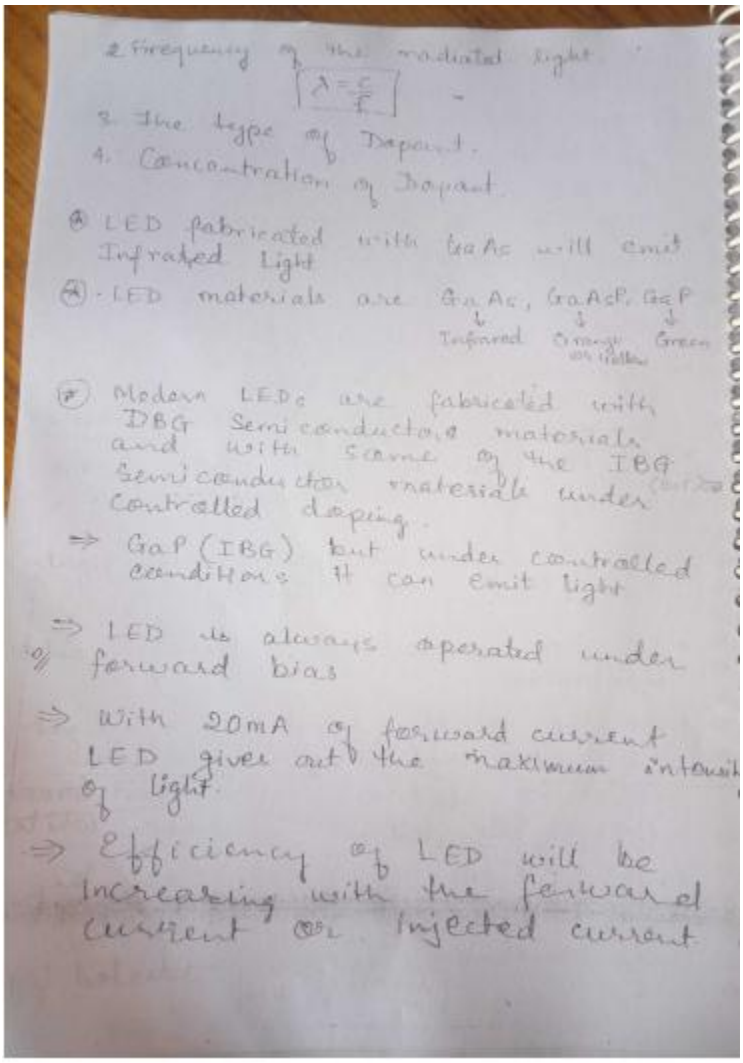
=> In LED light is emitted due to a large no of recombinations at the junction.

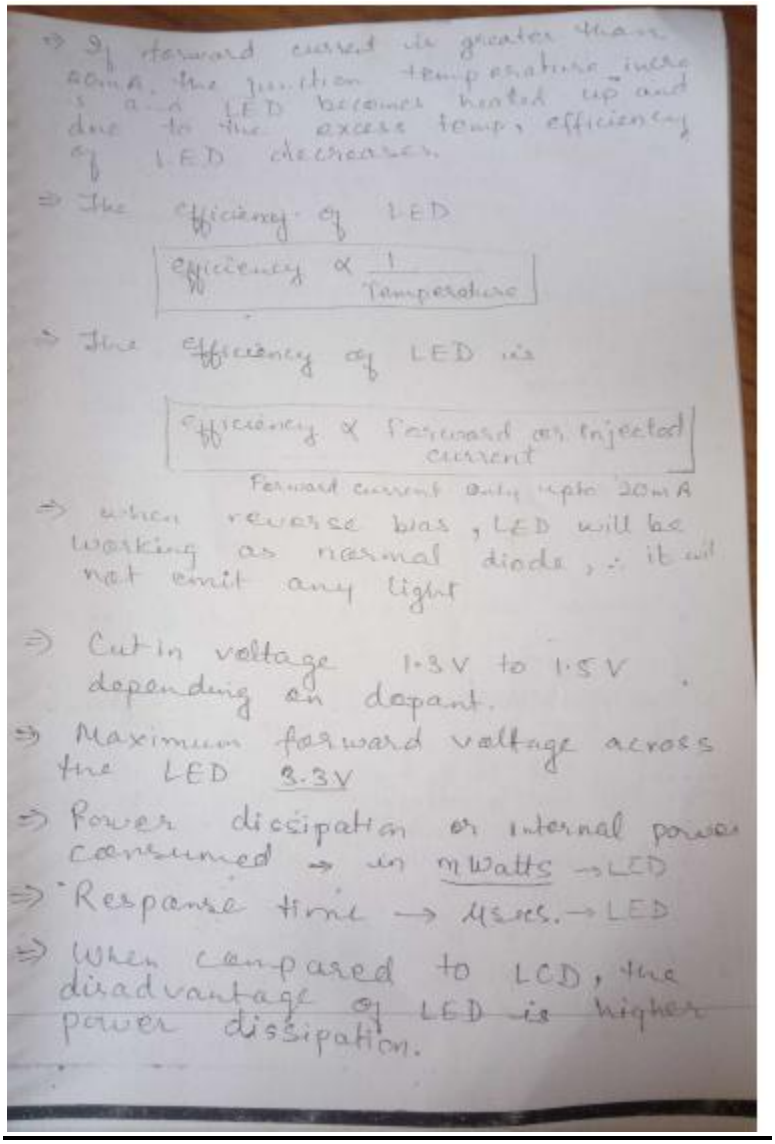
=> The best electro-luminescence device is LED

=> Generally fabricated with DBG Semiconductor, ~~property~~ most used material is GaAs.

- LED can emit the light either in the visible or in the invisible spectrum of light.
- In the invisible spectrum of light LED emits infrared light
- IR LED is used as a remote control transmitter.
- In the visible spectrum of light LED can emit anyone of the following colours. like Red, Green, Orange, ~~Or~~ Yellow and Amber. white
- The colour of light given by LED depends on.
  1. Wavelength of the radiated light

$$\lambda = 1.24/E_g \text{ } \mu m$$


2. Frequency of the radiated light
$\lambda = \frac{c}{f}$
3. The type of Dopant.
4. Concentration of Dopant
(A) LED fabricated with GaAs will emit Infrared Light
(B) - LED materials are GaAs, GaAsP, GaP
Infrared
Orange or Yellow
Green
(C) Modern LEDs are fabricated with IBG Semiconductor materials and with some of the IBG Semiconductor materials under Controlled doping.
⇒ GaP (IBG) but under controlled conditions it can emit light
⇒ LED is always operated under forward bias
⇒ With 20mA of forward current LED give out the maximum intensity of light.
⇒ Efficiency of LED will be increasing with the forward current or Injected current



⇒ If forward current is greater than 20mA, the junction temperature increases and LED becomes heated up and due to the excess temp, efficiency of LED decreases.

⇒ The efficiency of LED

$$\text{efficiency} \propto \frac{1}{\text{Temperature}}$$

⇒ The efficiency of LED is

$$\text{efficiency} \propto \text{Forward or injected current}$$

Forward current only upto 20mA

⇒ when reverse bias, LED will be working as normal diode, ∴ it will not emit any light

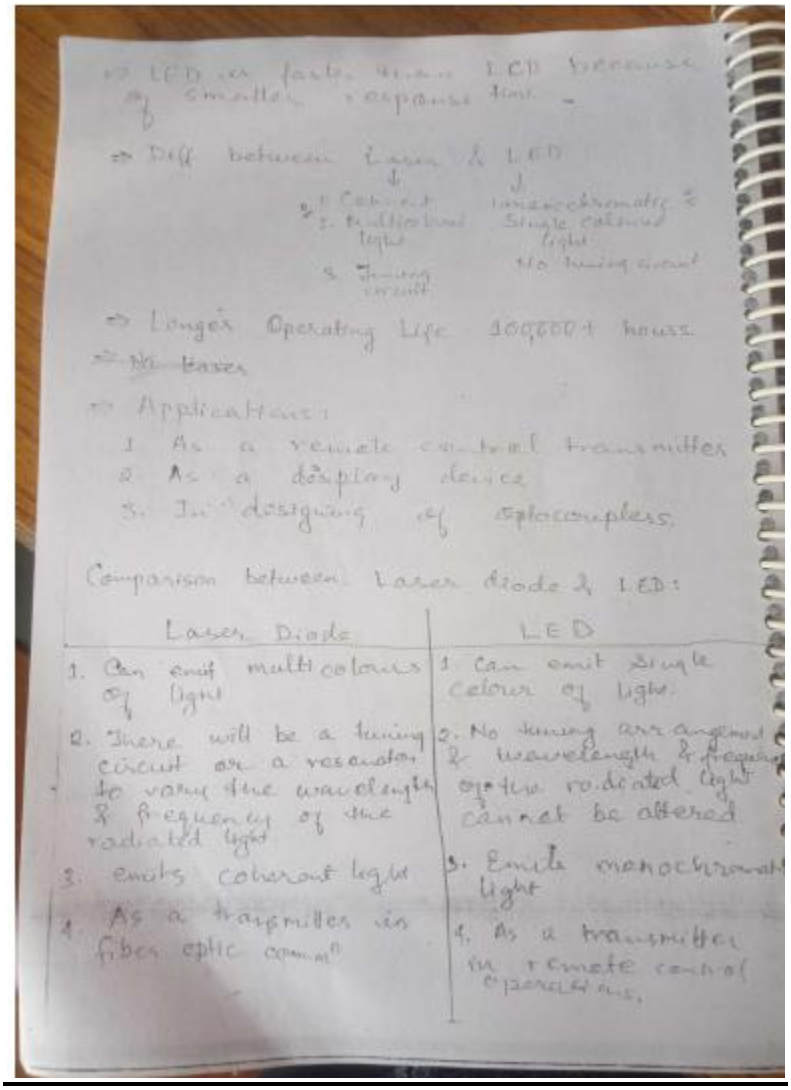
⇒ Cut in voltage 1.3V to 1.5V depending on dopant.

⇒ Maximum forward voltage across the LED 3.3V

⇒ Power dissipation or internal power consumed → in mWatts → LED

⇒ Response time → μsecs. → LED

⇒ When compared to LCD, the disadvantage of LED is higher power dissipation.



⇒ LED is faster than LCD because of smaller response time.

⇒ Diff between Laser & LED

Laser	LED
1. Coherent light	Incoherent & monochromatic light
2. Multicolored light	Single colored light
3. Tuning circuit	No tuning circuit

⇒ Longer Operating Life 100000+ hours.

⇒ ~~No~~ bases

⇒ Applications:
1. As a remote control transmitter
2. As a display device
3. In designing of optocouplers.

Comparison between Laser diode & LED:

Laser Diode	LED
1. Can emit multicolours of light	1. Can emit single colour of light.
2. There will be a tuning circuit or a resonator to vary the wavelength & frequency of the radiated light	2. No tuning arrangement & wavelength & frequency of the radiated light cannot be altered
3. emits coherent light	3. Emits monochromatic light
4. As a transmitter in fibre optic comm^n	4. As a transmitter in remote control operated ckts.

Photodiodes

Photodiodes – Basic principles

A photodiode is a type of photo detector capable of converting light into either current or voltage, depending upon the mode of operation.[1] The common, traditional solar cell used to generate electric solar power is a large area photodiode.

Photodiodes are similar to regular semiconductor diodes except that they may be either exposed (to detect vacuum UV or X-rays) or packaged with a window or optical fiber connection to allow light to reach the sensitive part of the device. Many diodes designed for use specifically as a photodiode use a PIN junction rather than a p-n junction, to increase the speed of response. A photodiode is designed to operate in reverse bias.

Fig-4

Principle of operation

A photodiode is a p-n junction or PIN structure. When a photon of sufficient energy strikes the diode, it excites an electron, thereby creating a free electron (and a positively charged electron hole). This mechanism is also known as the inner photoelectric effect. If the absorption occurs in the junction's depletion region, or one diffusion length away from it, these carriers are swept from the junction by the built-in field of the depletion region. Thus holes move toward the anode, and electrons toward the cathode, and a photocurrent is produced. This photocurrent is the sum of both the dark current (without light) and the light current, so the dark current must be minimized to enhance the sensitivity of the device.

Photovoltaic mode

When used in zero bias or photovoltaic mode, the flow of photocurrent out of the device is restricted and a voltage builds up. This mode exploits the photovoltaic effect, which is the basis for solar cells – a traditional solar cell is just a large area photodiode.

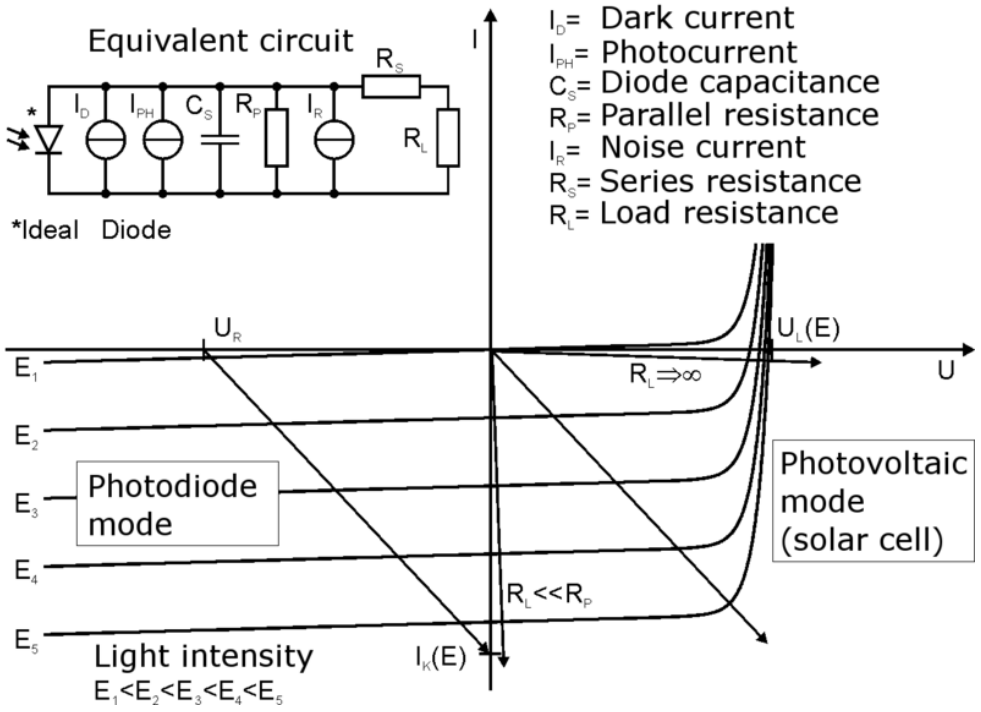


Fig-5

Photoconductive mode

In this mode the diode is often reverse biased (with the cathode positive), dramatically reducing the response time at the expense of increased noise. This increases the width of the depletion layer, which decreases the junction's capacitance resulting in faster response times. The reverse bias induces only a small amount of current (known as saturation or back current) along its direction while the photocurrent remains virtually the same. For a given spectral distribution, the photocurrent is linearly proportional to the luminance (and to the irradiance).

Although this mode is faster, the photoconductive mode tends to exhibit more electronic noise.[citation needed] The leakage current of a good PIN diode is so low (<1 nA) that the Johnson–Nyquist noise of the load resistance in a typical circuit often dominates.

Other modes of operation

Avalanche photodiodes have a similar structure to regular photodiodes, but they are operated with much higher reverse bias. This allows each photo-generated carrier to be multiplied by avalanche breakdown, resulting in internal gain within the photodiode, which increases the effective responsivity of the device.

A phototransistor is in essence a bipolar transistor encased in a transparent case so that light can reach the base-collector junction. The electrons that are generated by photons in the base-collector junction are injected into the base, and this photodiode current is amplified by the transistor's current gain β (or h_{fe}). If the emitter is left unconnected, the phototransistor becomes a photodiode. While phototransistors have a higher responsivity for light they are not able to detect low levels of light any better than photodiodes. Phototransistors also have significantly longer response times.

Critical performance parameters of a photodiode include:

Responsivity

The ratio of generated photocurrent to incident light power, typically expressed in A/W when used in photoconductive mode. The responsivity may also be expressed as a Quantum efficiency, or the ratio of the number of photogenerated carriers to incident photons and thus a unit less quantity.

Dark current

The current through the photodiode in the absence of light, when it is operated in photoconductive mode. The dark current includes photocurrent generated by background radiation and the saturation current of the semiconductor junction. Dark current must be accounted for by calibration if a photodiode is used to make an accurate optical power measurement, and it is also a source of noise when a photodiode is used in an optical communication system.

Noise-equivalent power

(NEP) The minimum input optical power to generate photocurrent, equal to the rms noise current in a 1 hertz bandwidth. The related characteristic detectivity (D) is the inverse of NEP, 1/NEP; and the specific detectivity ($D^{\star}$) is the detectivity normalized to the area (A) of the photodetector, $D^{\star} = D\sqrt{A}$. The NEP is roughly the minimum detectable input power of a photodiode.

When a photodiode is used in an optical communication system, these parameters contribute to the sensitivity of the optical receiver, which is the minimum input power required for the receiver to achieve a specified bit error rate.

Applications

P-N photodiodes are used in similar applications to other photodetectors, such as photoconductors, charge-coupled devices, and photomultiplier tubes. They may be used to generate an output which is dependent upon the illumination (analog; for measurement and the like), or to change the state of circuitry (digital; either for control and switching, or digital signal processing).

Photodiodes are used in consumer electronics devices such as compact disc players, smoke detectors, and the receivers for infrared remote control devices used to control equipment from televisions to air conditioners. For many applications either photodiodes or photoconductors may be used. Either type of photosensor may be used for light measurement, as in camera light meters, or to respond to light levels, as in switching on street lighting after dark.

Photosensors of all types may be used to respond to incident light, or to a source of light which is part of the same circuit or system. A photodiode is often combined into a single component with an emitter of light, usually a light-emitting diode (LED), either to detect the presence of a mechanical obstruction to the beam (slotted optical switch), or to couple two digital or analog circuits while maintaining extremely high electrical isolation between them, often for safety (optocoupler).

Photodiodes are often used for accurate measurement of light intensity in science and industry. They generally have a more linear response than photoconductors.

They are also widely used in various medical applications, such as detectors for computed tomography (coupled with scintillators), instruments to analyze samples (immunoassay), and pulse oximeters.

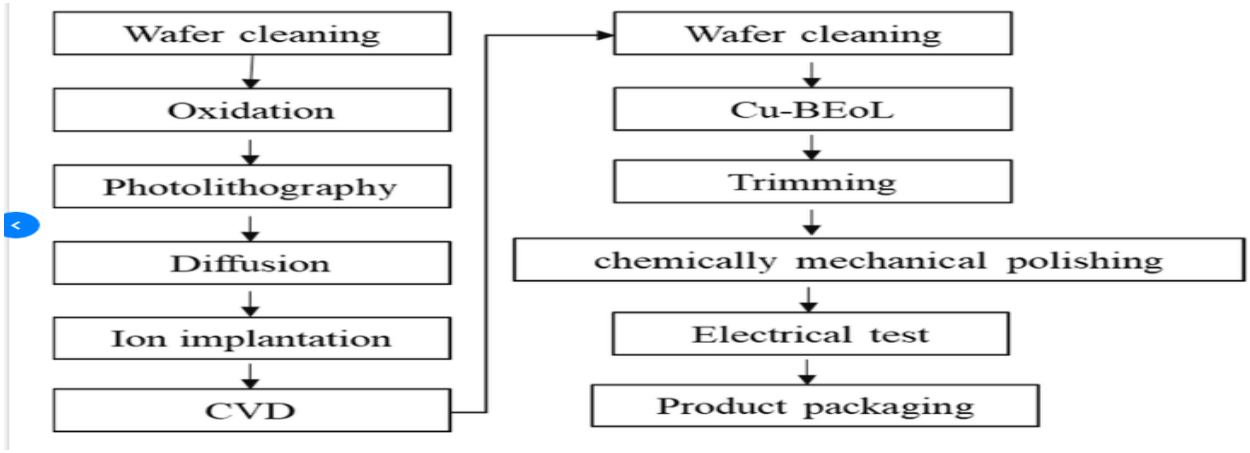
PIN diodes are much faster and more sensitive than p-n junction diodes, and hence are often used for optical communications and in lighting regulation.

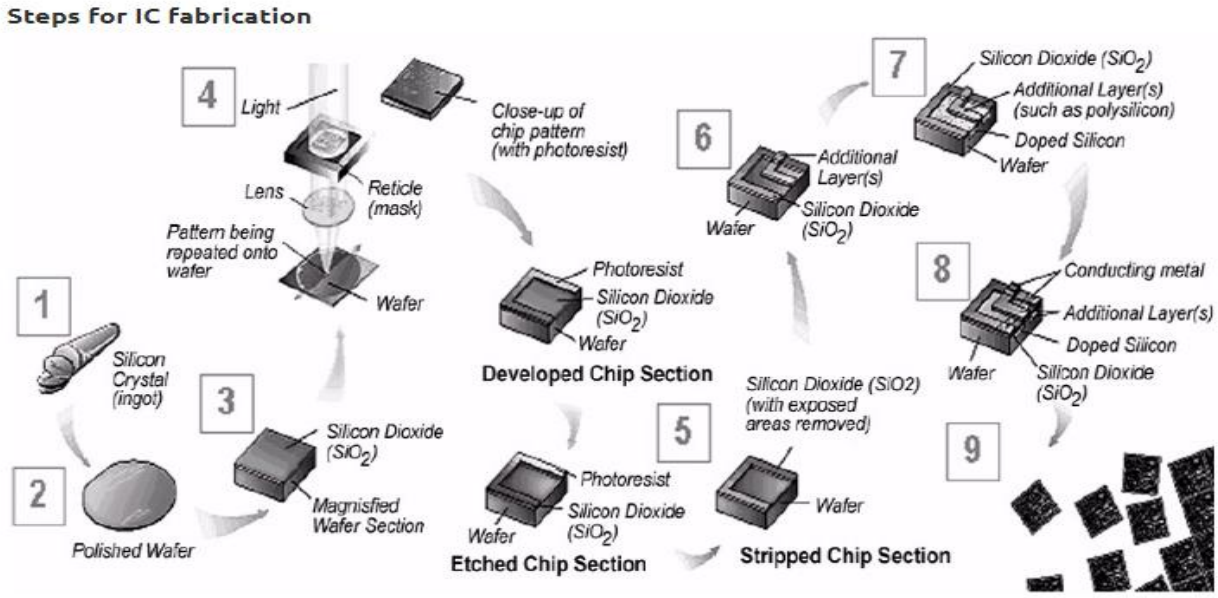
P-N photodiodes are not used to measure extremely low light intensities. Instead, if high sensitivity is needed, avalanche photodiodes, intensified charge-coupled devices or photomultiplier tubes are used for applications such as astronomy, spectroscopy, night vision equipment and laser rangefinding.

Integrated Circuit (IC)

An Integrated Circuit (IC) is also called as chip or microchip. It is a semiconductor wafer in which millions of components are fabricated. The active and passive components such as resistors, diodes, transistors etc and external connections are usually fabricated in on extremely tiny single chip of silicon. All circuit components and interconnections are formed on single thin wafer (substrate) is called monolithic IC. IC is very small in size. It require microscope to see connections between components. The steps to fabricate IC chips is similar to the steps required to fabricate transistors, diodes etc. In IC chips, the fabrication of circuit elements such as **transistors, diodes**, **capacitors** etc. and their interconnections are done at same time. It has so many advantages such as extremely small size, small weight, low cost, low power consumption, .high processing speed, easy replacement, etc. IC is the principal component in all electronic devices. **IC** can function as **amplifier**, **oscillator**, **timer**, counter, computer **memory** etc.

The starting material for integrated circuit (IC) fabrication is the single crystal silicon wafer. The end product of fabrication is functioning chips that are ready for packaging and electrical testing before being shipped to the customer. The intermediate steps are referred to as wafer fabrication (in- cluding sort). Wafer fabrication refers to the set of manufacturing processes
used to create semiconductor devices and circuits. Some common wafer terminology used are chip, die, device, circuit, and microchip. These refer to patterns covering the wafer surface that provide speci_c functionality. The terminology die and chip are most commonly used and interchangeably refer to one standalone unit on the wafer surface. Thus, a wafer can be said to be divided into many dies or chips.





The manufacturing of Integrated Circuits (IC) consists of following steps. The steps include 8-20 patterned layers created into the substrate to form the complete integrated circuit. The electrically active regions are created due to this layering in and on the surface of wafer. Hundreds of integrated circuits can be made on single thin silicon. Then it is cut into individual IC chips.

Step 1

Wafer production

The first step is wafer production. The wafer is a round slice of semiconductor material such as silicon. Silicon is preferred due to its characteristics. It is more suitable for manufacturing IC. It is the base or substrate for entire chip. First purified polycrystalline silicon is created from the sand. Then it is heated to produce molten liquid. A small piece of solid silicon is dipped on the molten liquid. Then the solid silicon (seed) is slowly pulled from the melt. The liquid cools to form single crystal ingot. A thin round wafer of silicon is cut using wafer slicer. Wafer slicer is a precise cutting machine and each slice having thickness about .01 to .025inches. When wafer is sliced, the surface will be damaged. It can be smoothening by polishing. After polishing the wafer, it must thoroughly clean and dried. The wafers are cleaned using high purity low particle chemicals .The silicon wafers are exposed to ultra pure oxygen.

Epitaxial growth

It means the growing of single silicon crystal upon original silicon substrate. A uniform layer of silicon dioxide is formed on the surface of wafer.

Step 2

Masking

To protect some area of wafer when working on another area, a process called **photolithography** is used. The process of photolithography includes masking with a photographic mask and photo etching. A photoresist film is applied on the wafer. The wafer is aligned to a mask using photo aligner. Then it is exposed to ultraviolet light through mask. Before that the wafer must be aligned with the mask. Generally, there are automatic tools for alignment purpose.

Step 3

Etching

It removes material selectively from the surface of wafer to create patterns. The pattern is defined by etching mask. The parts of material are protected by this etching mask. Either wet (chemical) or dry (physical) etching can be used to remove the unmasked material. To perform **etching** in all directions at same time, isotropic etching will be used. Anisotropic etching is faster in one direction. Wet etching is isotropic, but the etching time control is difficult. Wet etching uses liquid solvents for removing materials. It is not suited to transfer pattern with submicron feature size. It does not damage the material. Dry etching uses gases to remove materials. It is strongly anisotropic. But it is less selective. It is suited to transfer pattern having small size. The remaining photo resist is finally removed using additional chemicals or plasma. Then the wafer is inspected to make sure that the image is transferred from mask to the top layer of wafer.

Step 4

Doping

To alter the electrical character of silicon, atom with one less electron than silicon such as boron and atom with one electron greater then silicon such as phosphorous are introduced into the area. The P-type (boron) and N-type (phosphorous) are created to reflect their conducting characteristics. Diffusion is defined as the movement of impurity atoms in semiconductor material at high temperature.

Atomic diffusion

In this method p and n regions are created by adding dopants into the wafer. The wafers are placed in an oven which is made up of quartz and it is surrounded with heating elements. Then the wafers are heated at a temperature of about 1500-2200°F. The inert gas carries the dopant chemical. The dopant and gas is passed through the wafers and finally the dopant will get deposited on the wafer. This method can only be used for large areas. For small areas it will be difficult and it may not be accurate.

Ion implantation

This is also a method used for adding dopants. In this method, dopant gas such as phosphine or boron trichloride will be ionized first. Then it provides a beam of high energy dopant ions to the specified regions of wafer. It will penetrate the wafer. The depth of the penetration depends on the energy of the beam. By altering the beam energy, it is possible to control the depth of penetration of dopants into the wafer. The beam current and time of exposure is used to control the amount of dopant. This method is slower than atomic diffusion process. It does not require masking and this process is very precise. First it points the wafer that where it is needed and shoot the dopants to the place where it is required.

Step 5

Metallization

It is used to create contact with silicon and to make interconnections on chip. A thin layer of aluminum is deposited over the whole wafer. Aluminum is selected because it is a good conductor, has good mechanical bond with silicon, forms low resistance contact and it can be applied and patterned with single deposition and etching process.

Making successive layers: - The process such as masking, etching, doping will be repeated for each successive layers until all integrated chips are completed. Between the components, silicon dioxide is used as insulator. This process is called chemical vapor deposition. To make contact pads, aluminum is deposited. The fabrication includes more than three layers separated by dielectric layers. For electrical and physical isolation a layer of solid dielectric is surrounded in each component which provides isolation. It is possible to fabricate PNP and NPN transistor in the same silicon substrate. To avoid damage and contamination of circuit, final dielectric layer (passivation) is deposited. After that, the individual IC will be tested for electrical function. Check the functionality of each chip on wafer. Those chips are not passed in the test will be rejected.

Assembly and packaging

Each of the wafers contains hundreds of chips. These chips are separated and packaged by a method called scribing and cleaving. The wafer is similar to a piece of glass. A diamond saw cut the wafer into single chips. The diamond tipped tool is used to cut the lines through the rectangular grid which separates the individual chips. The chips that are failed in electrical test are discarded. Before packaging, remaining chips are observed under microscope. The good chip is then mounted into a package. Thin wire is connected using ultrasonic bonding. It is then encapsulated for protection. Before delivered to customer, the chip is tested again. There are three configurations available for packaging. They are metal can package, ceramic flat package and dual in line package. For military applications, the chip is assembled in ceramic packages. The complete integrated circuits are sealed in anti static plastic bags.

Duel-well Process or Twin-tub Process :

In Duel-well process both p-well and n-well for NMOS and PMOS transistors respectively are formed on the same substrate. The main advantage of this process is that the threshold voltage, body effect parameter and the transconductance can be optimized separately. The starting material for this process is p+ substrate with epitaxially grown p-layer which is also called as epilayer. The process steps of twin-tub process are shown in Figure below.

The process starts with a p-substrate surfaced with a lightly doped p-epitaxial layer.

Step 1 : A thin layer of SiO_2 is deposited which will serve as the pad oxide.

Step 2 : A thicker sacrificial silicon nitride layer is deposited by chemical vapour deposition.

Step 3 : A plasma etching process is used to create trenches used for insulating the devices.

Step 4 : The trenches are filled with SiO_2 which is called as the field oxide.

Step 5 : To provide flat surface chemical mechanical planarization is performed and also sacrificial nitride and pad oxide is removed.

Step 6 : The p-well mask is used to expose only the p-well areas, after this implant and annealing sequence is applied to adjust the well doping. This is followed by second implant step to adjust the threshold NMOS transistor.

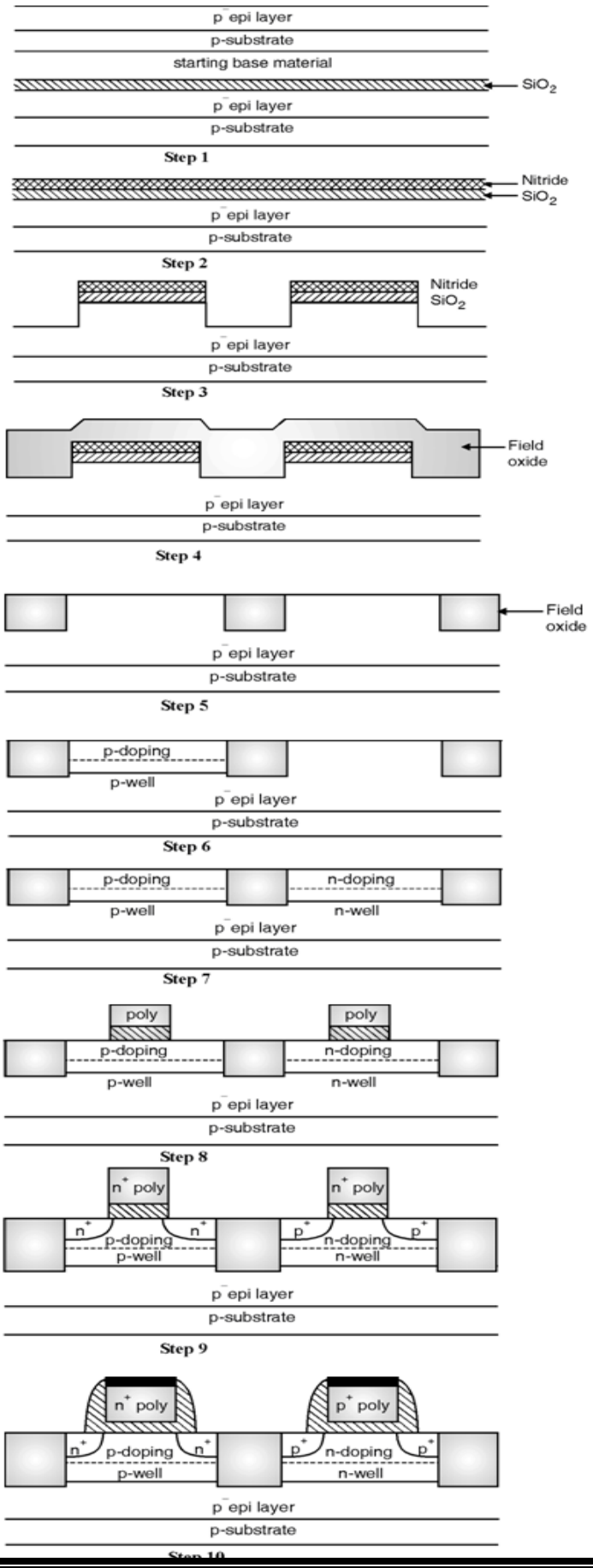
Step 7 : The n-well mask is used to expose only the n-well areas, after this implant and annealing sequence is applied to adjust the well doping. This is followed by a second implant step to adjust the threshold voltage of PMOS transistor.

Step 8 : A thin layer of gate oxide and polysilicon is chemically deposited and patterned with the help of polysilicon mask.

Step 9 : Ion implantation to dope the source and drain regions of the PMOS (p^+) and NMOS (n^+) transistors is used this will also form n^+ polysilicon gate and p^+ polysilicon gate for NMOS and PMOS transistors respectively.

Step 10 : Then the oxide or nitride spacers are formed by chemical vapour deposition (CVD).

Step 11 : In this step contact or holes are etched, metal is deposited and patterned. After the deposition of last metal layer final passivation or overglass is deposited for protection.



Dual well process steps

The Fabrication Process of CMOS Transistor

There was an era, where computers were such mammoth in size that to install them, easily a room space was required. But today they are so evolved that we can even carry them as notebooks easily. The innovation that made this possible was the concept of Integrated Circuits. In **Integrated Circuits**, a large number of active and **passive elements** along with their interconnections are developed over a small silicon wafer typically of 50 by 50 mils in cross section. The basic processes followed for production of such circuits include epitaxial growth, masked impurity diffusion, oxide growth, and oxide etching, using photolithography for making pattern.
The components over the wafer include resistors, transistors, diodes, capacitors etc… The most complicated element to manufacture over IC's is transistors. **Transistors are of various types** such as CMOS, BJT, FET. We choose the type of transistor technology to be implemented over an IC based on requirements. In this article let us get familiarized with the concept of **CMOS fabrication** (or) fabrication of transistors as CMOS.

CMOS Fabrication

For less power dissipation requirement **CMOS technology** is used for implementing transistors. If we require a faster circuit then transistors are implemented over **IC using BJT**. Fabrication of **CMOS transistors** as IC's can be done in three different methods.
The N-well / P-well technology, where n-type diffusion is done over a p-type substrate or p-type diffusion is done over n-type substrate respectively.

The **Twin well technology**, where **NMOS and PMOS transistor** are developed over the wafer by simultaneous diffusion over an epitaxial growth base, rather than a substrate.
The silicon On Insulator process, where rather than using silicon as the substrate an insulator material is used to improve speed and latch-up susceptibility.

N- well/ P- well Technology

CMOS can be obtained by integrating both **NMOS and PMOS transistors** over the same silicon wafer. In N–well technology an n-type well is diffused on a p-type substrate whereas in P- well it is vice- verse.

CMOS Fabrication Steps

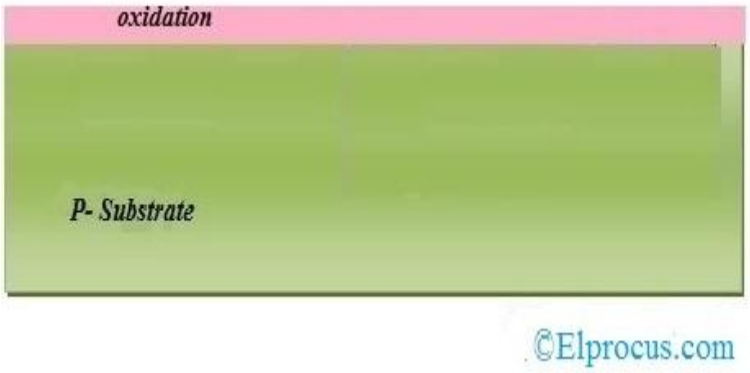
The **CMOS fabrication process flow** is conducted using twenty basic fabrication steps while manufactured using N- well/P-well technology.

Making of CMOS using N well

Step 1: First we choose a substrate as a base for fabrication. For N- well, a P-type silicon substrate is selected.

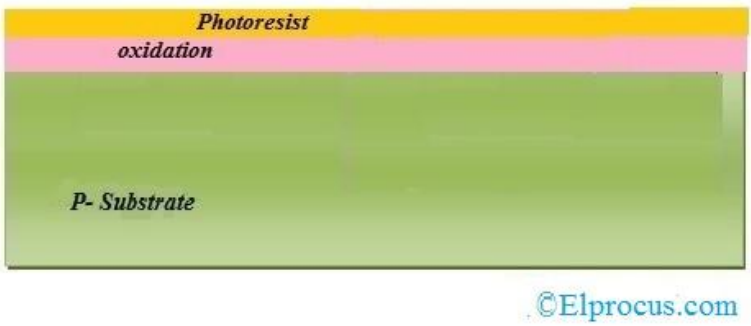
Substrate

Step 2 – Oxidation: The selective diffusion of n-type impurities is accomplished using SiO2 as a barrier which protects portions of the wafer against contamination of the substrate. SiO_2 is laid out by oxidation process done exposing the substrate to high-quality oxygen and hydrogen in an oxidation chamber at approximately 1000^0c



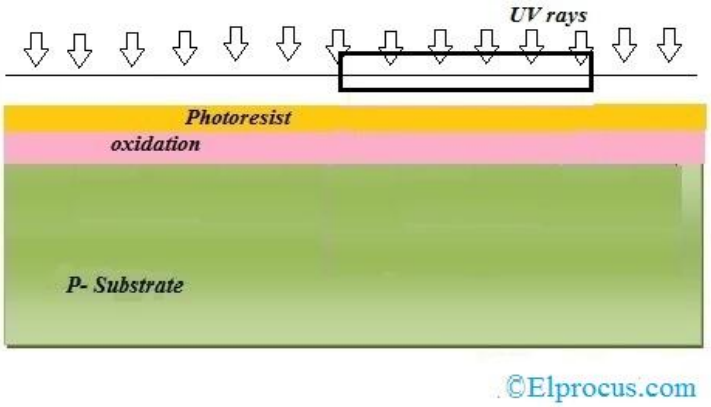
Oxidation

Step 3 – Growing of Photoresist: At this stage to permit the selective etching, the SiO2 layer is subjected to the photolithography process. In this process, the wafer is coated with a uniform film of a photosensitive emulsion.



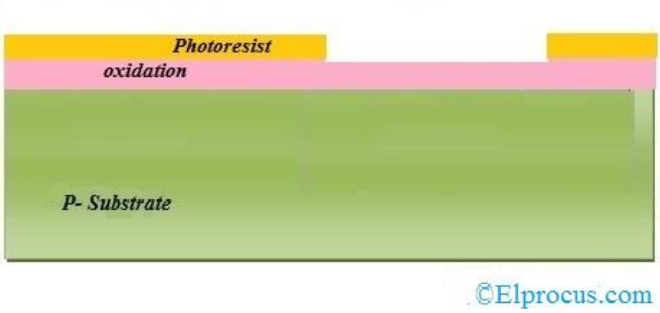
Growing of Photoresist

Step 4 – Masking: This step is the continuation of the photolithography process. In this step, a desired pattern of openness is made using a stencil. This stencil is used as a mask over the photoresist. The substrate is now exposed to **UV rays** the photoresist present under the exposed regions of mask gets polymerized.



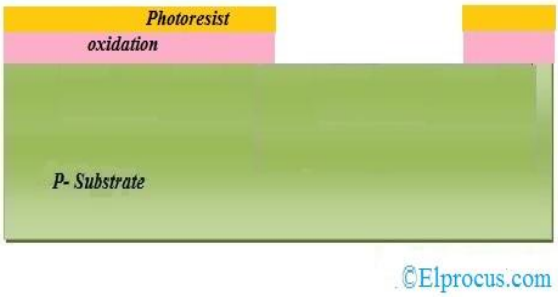
Masking of Photoresist

Step 5 – Removal of Unexposed Photoresist: The mask is removed and the unexposed region of photoresist is dissolved by developing wafer using a chemical such as Trichloroethylene.



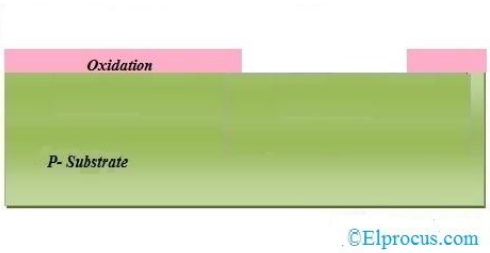
Removal of Photoresist

Step 6 – Etching: The wafer is immersed in an etching solution of hydrofluoric acid, which removes the oxide from the areas through which dopants are to be diffused.



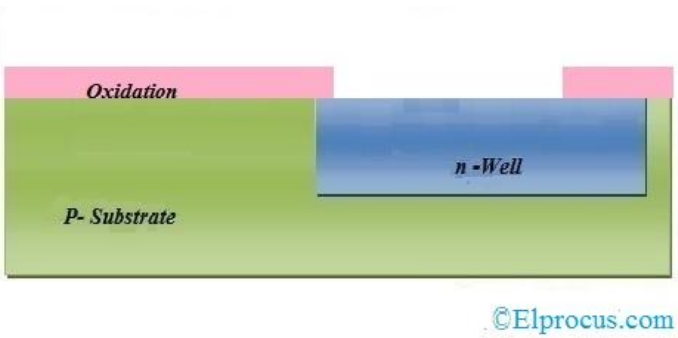
Etching of SiO2

Step 7 – Removal of Whole Photoresist Layer: During the **etching process**, those portions of SiO2 which are protected by the photoresist layer are not affected. The photoresist mask is now stripped off with a chemical solvent (hot H2SO4).



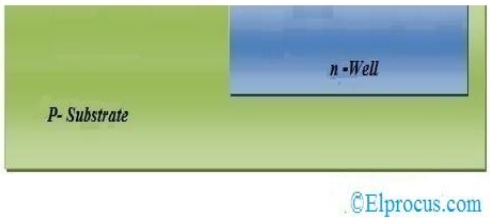
Removal of Photoresist Layer

Step 8 – Formation of N-well: The n-type impurities are diffused into the p-type substrate through the exposed region thus forming an N- well.



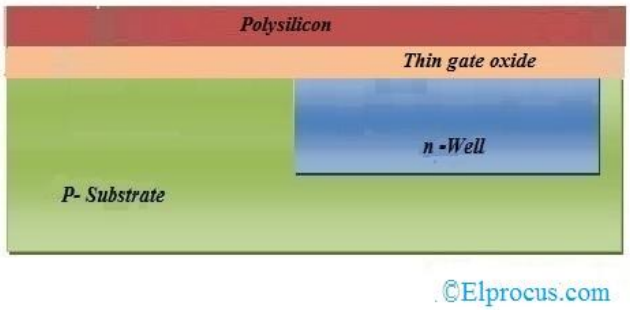
Formation of N-well

Step 9 – Removal of SiO2: The layer of SiO2 is now removed by using hydrofluoric acid.



Removal of SiO2

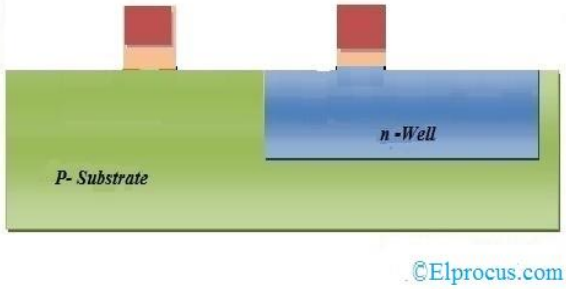
Step 10 – Deposition of Polysilicon: The misalignment of the gate of a **CMOS transistor** would lead to the unwanted capacitance which could harm circuit. So to prevent this "Self-aligned gate process" is preferred where gate regions are formed before the formation of source and drain using ion implantation.



Deposition of Polysilicon

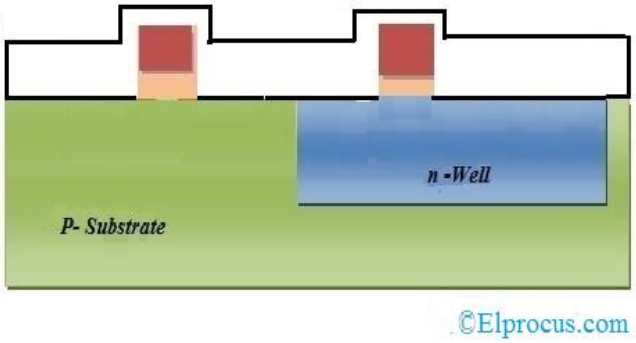
Polysilicon is used for formation of the gate because it can withstand the high temperature greater than 8000^0c when a wafer is subjected to annealing methods for formation of source and drain. Polysilicon is deposited by using **Chemical Deposition Process** over a thin layer of gate oxide. This thin gate oxide under the Polysilicon layer prevents further doping under the gate region.

Step 11 – Formation of Gate Region: Except the two regions required for formation of the gate for **NMOS and PMOS transistors** the remaining portion of Polysilicon is stripped off.



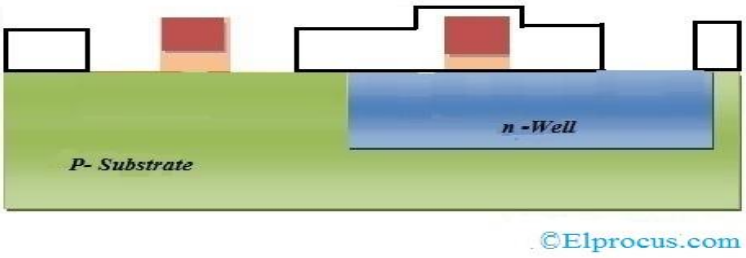
Formation of Gate Region

Step 12 – Oxidation Process: An oxidation layer is deposited over the wafer which acts as a shield for further **diffusion and metallization processes**.



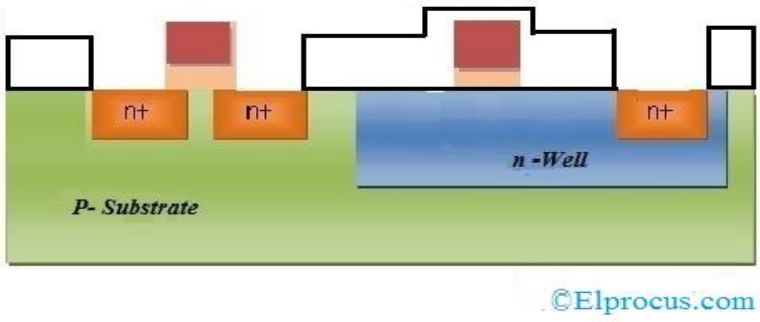
Oxidation Process

Step 13 – Masking and Diffusion: For making regions for diffusion of n-type impurities using masking process small gaps are made.



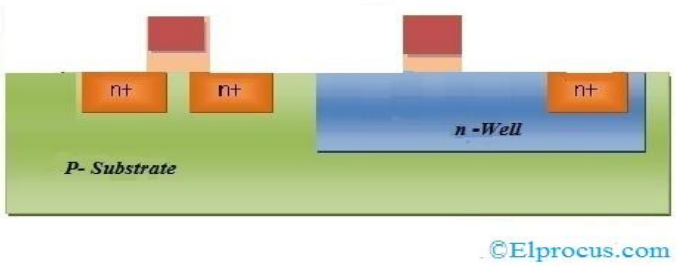
Masking

Using diffusion process three n+ regions are developed for the formation of terminals of NMOS.



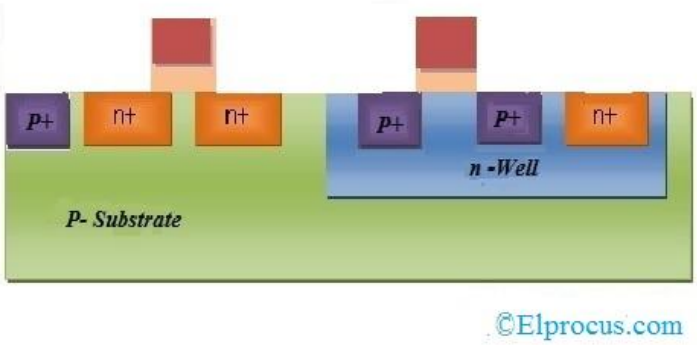
N-diffusion

Step 14 – Removal of Oxide: The oxide layer is stripped off.



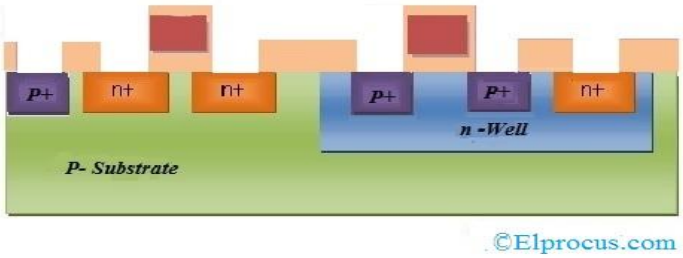
Removal of Oxide

Step 15 – P-type Diffusion: Similar to the n-type diffusion for forming the terminals of PMOS p-type diffusion are carried out.



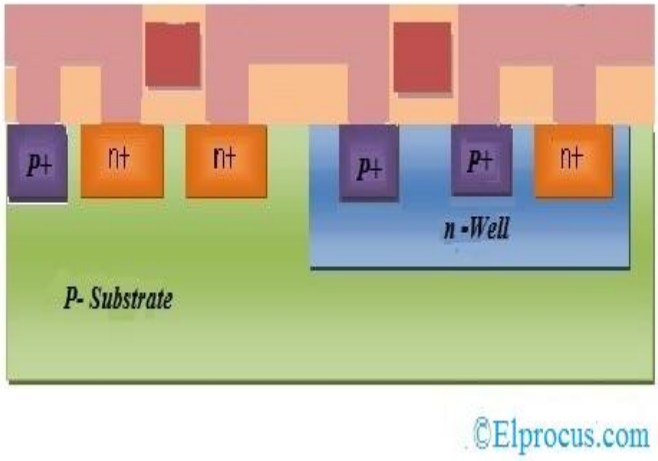
P-Type Diffusion

Step 16 – Laying of Thick Field oxide: Before forming the metal terminals a thick field oxide is laid out to form a protective layer for the regions of the wafer where no terminals are required.



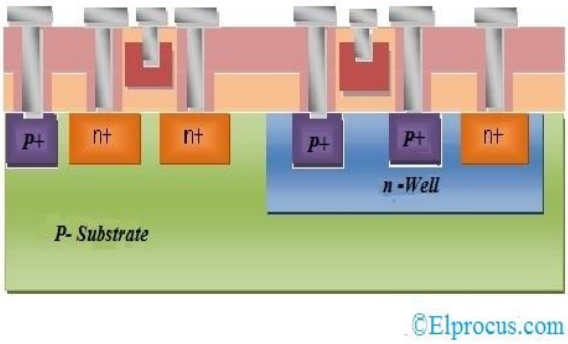
Thick Field oxide Layer

Step 17 – Metallization: This step is used for the formation of metal terminals which can provide interconnections. Aluminum is spread on the whole wafer.



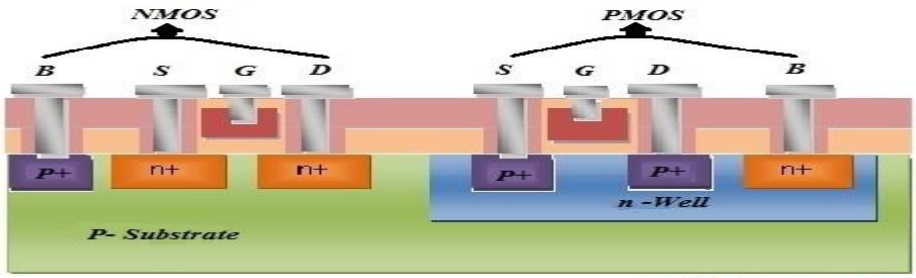
Metallization

Step 18 – Removal of Excess Metal: The excess metal is removed from the wafer.
Step 19 – Formation of Terminals: In the gaps formed after removal of excess metal terminals are formed for the interconnections.



Formation of Terminals

Step 20 – Assigning the Terminal Names: Names are assigned to the terminals of **NMOS and PMOS transistors**.



Assigning Terminal names

Making of CMOS using P well Technology

The p-well process is similar to N well process except that here n-type substrate is used and p-type diffusions are carried out. For simplicity usually, N well process is preferred.

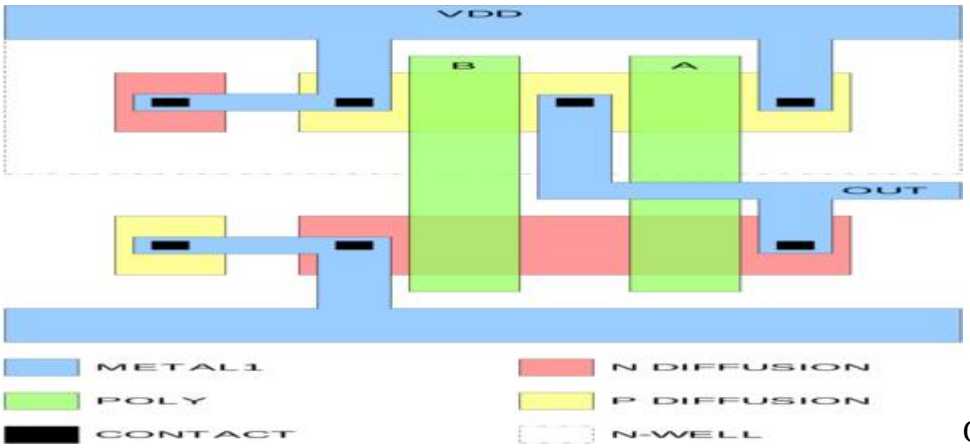
Twin Tube Fabrication of CMOS

Using Twin-tube process one can control the gain of P and N-type devices. Various steps involved in the **fabrication of CMOS using Twin-tube method** are as follows

- A lightly doped n or p-type substrate is taken and the epitaxial layer is used. Epitaxial layer protects the latch-up problem in the chip.
- The high purity silicon layers with measured thickness and exact dopant concentration are grown.
- Formation of tubes for P and N well.
- Thin oxide construction for protection from contamination during diffusion processes.
- Source and drain are formed using ion implantation methods.
- Cuts are made for making portions for metal contacts.
- Metallization is done for drawing metal contacts

CMOS IC Layout

The upper view of **a CMOS fabrication and layout** is given. Here various metal contacts and N well diffusions can be viewed clearly.



CMOS IC Layout

Thus, this is all about **CMOS fabrication techniques**. Let us consider a 1-in-square wafer divided into 400 chips of surface area 50 mil by 50 mils. It takes an area of 50 mil2 to fabricate a transistor. Hence each IC contains 2 transistors thus there are 2 x 400 = 800 transistors built on each wafer. If 10 wafers are processed each batch then 8000 transistors can be manufactured simultaneously. What are various components have you observed on an IC?

Manufacturing CMOS Integrated Circuits

Introduction

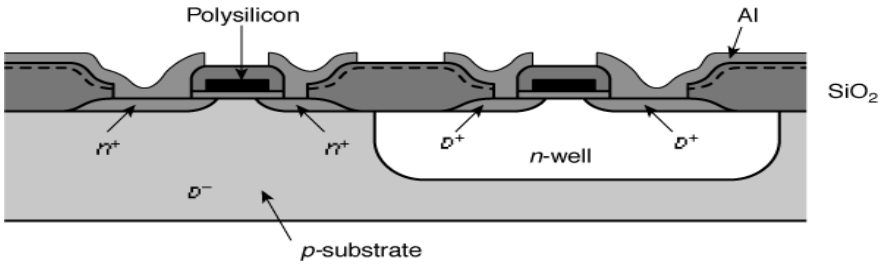
Most digital designers will never be confronted with the details of the manufacturing process that lies at the core of the semiconductor revolution. Yet, some insight in the steps that lead to an operational silicon chip comes in quite handy in understanding the physical constraints that are imposed on a designer of an integrated circuit, as well as the impact of the fabrication process on issues such as cost.

In this chapter, we briefly describe the steps and techniques used in a modern inte-

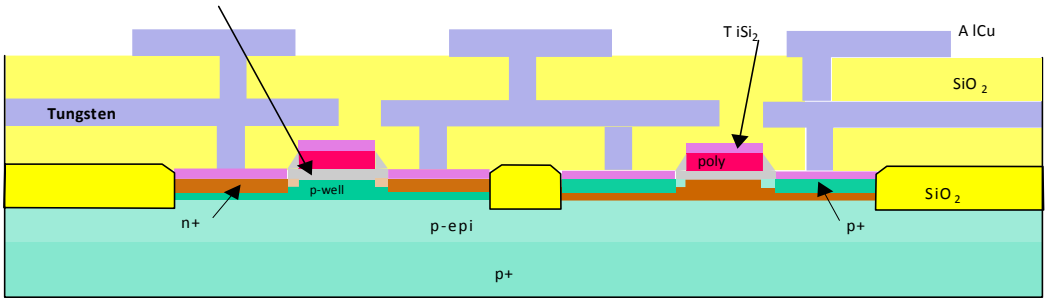
grated circuit manufacturing process. It is not our aim to present a detailed description of the fabrication technology, which easily deserves a complete course [Plummer00]. Rather we aim at presenting the general outline of the flow and the interaction between the vari- ous steps. We learn that a set of *optical masks* forms the central interface between the intrinsics of the manufacturing process and the design that the user wants to see trans- ferred to the silicon fabric. The masks define the patterns that, when transcribed onto the different layers of the semiconductor material, form the elements of the electronic devices and the interconnecting wires. As such, these patterns have to adhere to some constraints in terms of minimum width and separation if the resulting circuit is to be fully functional. This collection of constraints is called the *design rule set*, and acts as the contract between the circuit designer and the process engineer. If the designer adheres to these rules, he gets a guarantee that his circuit will be manufacturable. An overview of the common design rules, encountered in modern CMOS processes, will be given. Finally, an overview is given of the *IC packaging* options. The package forms the interface between the circuit implemented on the silicon die and the outside world, and as such has a major impact on the performance, reliability, longevity, and cost of the integrated circuit.

Manufacturing CMOS Integrated Circuits

A simplified cross section of a typical CMOS inverter is shown in Figure 2.1. The CMOS process requires that both *n*-channel (NMOS) and *p*-channel (PMOS) transistors be built in the same silicon material. To accomodate both types of devices, special regions called *wells* must be created in which the semiconductor material is opposite to the type of the channel. A PMOS transistor has to be created in either an *n*-type substrate or an *n*-well, while an NMOS device resides in either a *p*-type substrate or a *p*-well. The cross section



shown in Figure 2.1 features an *n*-well CMOS process, where the NMOS transistors are implemented in the *p*-doped substrate, and the PMOS devices are located in the *n*-well. Increasingly, modern processes are using a *dual-well* approach that uses both *n*- and *p*- wells, grown on top on a epitaxial layer, as shown in Figure 2.2. We will restrict the remainder of this discussion to the latter process (without loss of generality).



gate-oxide

Cross section of modern dual-well CMOS process.

The CMOS process requires a large number of steps, each of which consists of a sequence of basic operations. A number of these steps and/or operations are executed very repetitively in the course of the manufacturing process. Rather than diving directly into a description of the overall process flow, we first discuss the starting material followed by a detailed perspective on some of the most-often recurring operations.

The Silicon Wafer

The base material for the manufacturing process comes in the form of a single-crystalline, lightly doped *wafer*. These wafers have typical diameters between 4 and 12 inches (10 and 30 cm, respectively) and a thickness of at most 1 mm, and are obtained by cutting a single- crystal ingot into thin slices (Figure 2.3). A starting wafer of the p^--type might be doped around the levels of 2 □

10^{21} impurities/m^3. Often, th surface of the wafer is dope more heavily, and a singl cristal *epitax- ial layer* of th opposite type is grown ove

the surface before the wafers are handed to the processing com- pany. One important metric is the defect density of the base material. High defect densities lead to a larger fraction of non-functional circuits, and consequently an increase in cost of the final product.

Single-crystal ingot and sliced wafers

Photolithography

In each processing step, a certain area on the chip is masked out using the appropriate opti- cal mask so that a desired processing step can be selectively applied to the remaining regions. The processing step can be any of a wide range of tasks including oxidation, etch- ing, metal and polisilicon deposition, and ion implantation. The technique to accomplish this selective masking, called *photolithography*, is applied throughout the manufacturing process. Figure 2.4 gives a graphical overview of the different operations involved in a typical photolitographic process. The following steps can be identified:

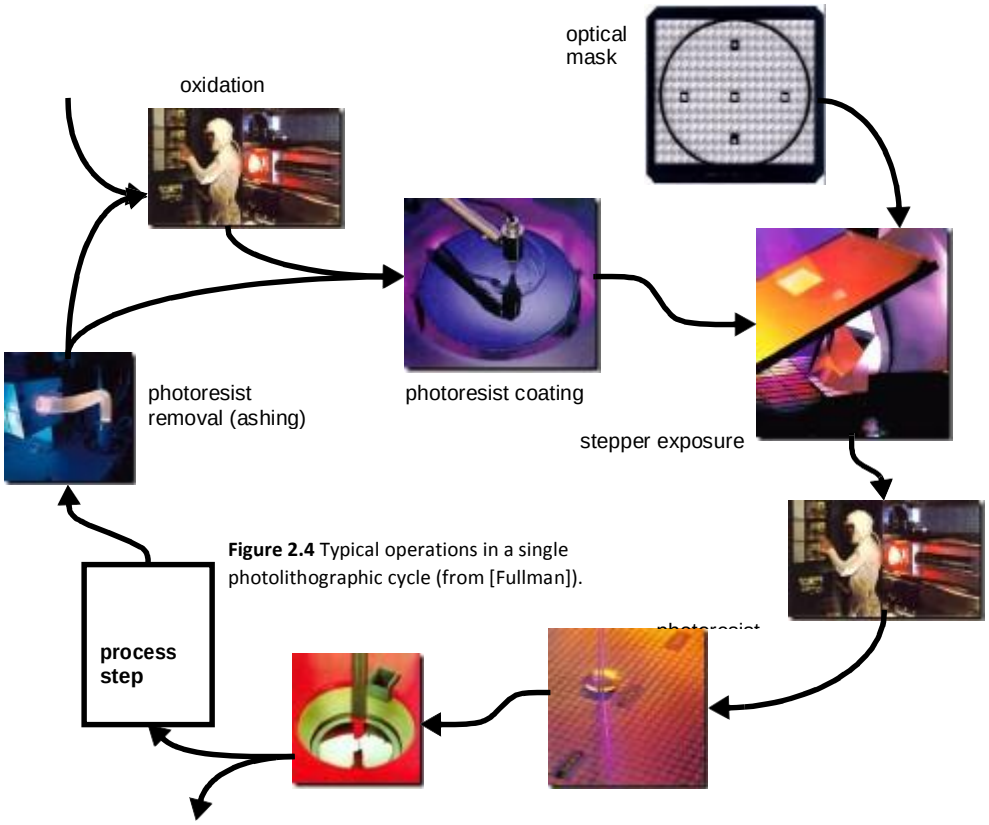


Figure 2.4 Typical operations in a single photolithographic cycle (from [Fullman]).

1. *Oxidation layering* — this optional step deposits a thin layer of SiO_2 over the com- plete wafer by exposing it to a mixture of high-purity oxygen and hydrogen at ± 1000□ C. The oxide is used as an insulation layer and also forms transistor gates.
2. *Photoresist coating* — a light-sensitive polymer (similar to latex) is evenly applied while

spinning the wafer to a thickness of approximately 1 □ m. This material is originallysoluble in an organic solvent, but has the property that the polymers cross- link when exposed to light, making the affected regions insoluble. A photoresist of
this type is called *negative*. A positive photoresist has the opposite properties; origi- nally insoluble, but soluble after exposure. By using both positive and negative resists, a single mask can sometimes be used for two steps, making complementary regions available for processing. Since the cost of a mask is increasing quite rapidly with the scaling of technology, a reduction of the number of masks is surely of high priority.

3. *Stepper exposure* — a glass mask (or reticle), containing the patterns that we want to transfer to the silicon, is brought in close proximity to the wafer. The mask is opaque in the regions that we want to process, and transparent in the others (assum- ing a negative photoresist). The glass mask can be thought of as the negative of one layer of the microcircuit. The combination of mask and wafer is now exposed to ultra-violet light. Where the mask is transparent, the photoresist becomes insoluble.

4. *Photoresist development and bake* — the wafers are developed in either an acid or base solution to remove the non-exposed areas of photoresist. Once the exposed photoresist is removed, the wafer is "soft-baked" at a low temperature to harden the remaining photoresist.

5. *Acid Etching* — material is selectively removed from areas of the wafer that are not covered by photoresist. This is accomplished through the use of many different types of acid, base and caustic solutions as a function of the material that is to be removed. Much of the work with chemicals takes place at large wet benches where special solutions are prepared for specific tasks. Because of the dangerous nature of some of these solvents, safety and environmental impact is a primary concern.

6. *Spin, rinse, and dry* — a special tools (called SRD) cleans the wafer with deionized water and dries it with nitogen. The microscopic scale of modern semiconductor devices means that even the smallest particle of dust or dirt can destroy the circuitry. To prevent this from happening, the processing steps are performed in ultra-clean rooms where the number of dust particles per cubic foot of air ranges between 1 and
10. Automatic wafer handling and robotics are used whenever possible. This explains why the cost of a state-of-the-art fabrication facility easily ranges in the multiple billions of dollars. Even then, the wafers must be constantly cleaned to avoid contamination, and to remove the left-over of the previous process steps.

7. *Various process steps* — the exposed area can now be subjected to a wide range of process steps, such as ion implantation, plasma etching, or metal deposition. These are the subjects of the subsequent section.

8. *Photoresist removal (or ashing)* — a high-temperature plasma is used to selectively remove the remaining photoresist without damaging device layers.

We illustrate the use of the photolitographic process for one specific example, the patterning

of a layer of SiO_2, in Figure 2.5. The sequence of process steps shown in the Figure patterns exactly one layer of the semiconductor material, and may seem very com- plex. Yet, the reader has to bear in mind that that same sequence patterns the layer of **the complete surface of the wafer**. It is hence a very parallel process, transferring hundreds of millions of patterns to the semiconductor surface simultaneously. The concurrent and

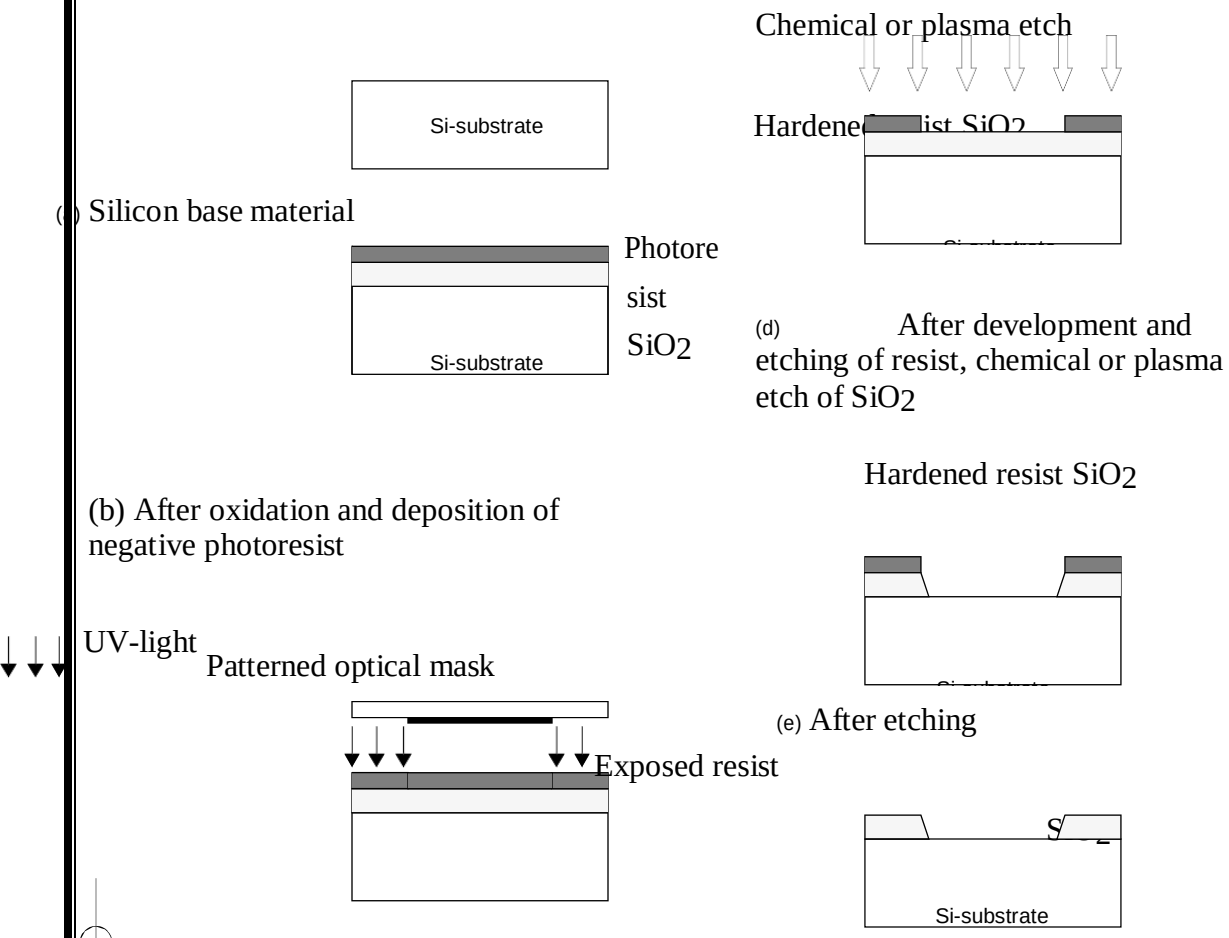


Figure 2.5 Process steps for patterning of SiO_2.

scalable nature of the optolithographical process is what makes the cheap manufacturing of complex semiconductor circuits possible, and lies at the core of the economic success of the semiconductor industry.

The continued scaling of the minimum feature sizes in integrated circuits puts an enormous

burden on the developer of semiconductor manufacturing equipment. This is especially true for the optolithographical process. The dimensions of the features to be transcribed approach the wavelengths of the optical light sources, so that achieving the necessary resolution and accuracy becomes harder and harder. So far, ingenous ingeneer- ing has extended the lifetime of this process at least until the 100 nm (or 0.1 □ m) process generation. Beyond that point, other solutions that offer a a finer resolution such as X-ray or electron-beam may be needed. These techniques, while fully functional, are currently less attractive from an economic viewpoint.

Some Recurring Process Steps Diffusion and Ion Implantation

Many steps of the integrated circuit manufacturing process require a chance in the dopant concentration of some parts of the material. The creation of the source and drain regions, well and substrate contacts, the doping of the polysilicon, and the adjustments of the device threshold are examples of such. There exist two approaches for introducing these dopants—diffusion and ion implantation. In both techniques, the area to doped is exposed, while the rest of the wafer is coated with a layer of buffer material, typically SiO_2.

In *diffusion implantation*, the wafers are placed in a quartz tube embedded in a

heated furnace. A gas containing the dopant is introduced in the tube. The high tempera- tures of the furnace, typically 900 to 1100 □ C, cause the dopants to diffuse into the exposed surface both vertically and horizontally. The final dopant concentration is the greatest at the surface and decreases in a gaussian profile deeper in the material.

In *ion implantation*, dopants are introduced as ions into the material. The ion implantation system directs and sweeps a beam of purified ions over the semiconductor surface. The acceleration of the ions determines how deep they will penetrate the material, while the beam current and the exposure time determine the dosage. The ion implantation method allows for an independent control of depth and dosage. This is the reason that ion implantation has largely displaced diffusion in modern semiconductor manufacturing.

Ion implantation has some unfortunate side effects however, the most important one being lattice damage. Nuclear collisions during the high energy implantation cause the displacement of substrate atoms, leading to material defects. This problem is largely resolved by applying a subsequent *annealing* step, in which the wafer is heated to around 1000□ C for 15 to 30 minutes, and then allowed to cool slowly. The heating step thermally vibrates the atoms, which allows the bonds to reform.

Deposition

Any CMOS process requires the repetitive deposition of layers of a material over the complete wafer, to either act as buffers for a processing step, or as insulating or con- ducting layers. We have already discussed the oxidation process, which allows a layer of SiO_2 to be grown. Other materials require different techniques. For instance, silicon nitride (Si_3N_4) is used as a sacrificial buffer material during the formation of the field oxide and the introduction of the stopper implants. This silicon nitride is deposited every- where using a process called *chemical vapor deposition* or CVD, which uses a gas-phase reaction with energy supplied by heat at around 850□ C.

Polysilicon, on the other hand, is deposited using a chemical deposition process, which flows silane gas over the heated wafer coated with SiO_2 at a temperature of approx- imately 650□ C. The resulting reaction produces a non-cristalline or amorphous material called polysilicon. To increase to conductivity of the material, the deposition has to be fol- lowed by an implantation step.

The Aluminum interconnect layers are typically deployed using a process known as *sputtering*. The aluminum is evaporated in a vacuum, with the heat for the evaporation

delivered by electron-beam or ion-beam bombarding. Other metalic interconnect materials such as Copper require different deposition techniques.

Etching

Once a material has been deposited, etching is used to selectively form patterns such as wires and contact holes. The *wet etching* process was described earlier, and makes use of acid acid or basic solutions. For instance, hydrofluoric acid buffered with ammonium fluoride is typically used to etch SiO_2.
In recent years, *dry* or *plasma etching* has made a lot of inroad. A wafer is placed

into the etch tool's processing chamber and given a negative electrical charge. The cham- ber is heated to 100□ C and brought to a vacuum level of 10 millitorrs, then filled with a positively charged plasma (usually a mix of nitrogen, chlorine and boron trichloride). The opposing electrical charges cause the rapidly moving plasma molecules to align them- selves in a vertical direction, forming a microscopic chemical and physical "sandblasting" action which removes the exposed material. Plasma etching has the advantage of offering a well-defined directionality to the etching action, creating patterns with sharp vertical contours.

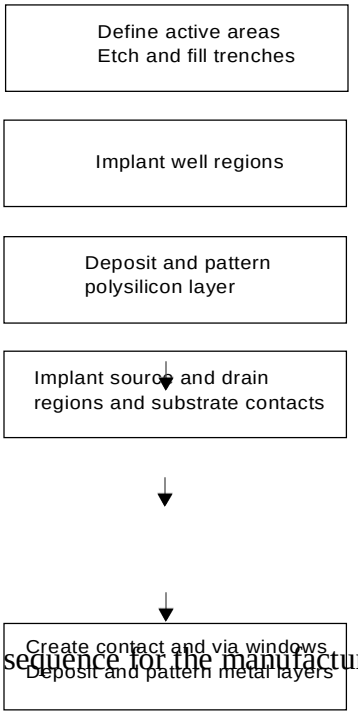
Planarization

To reliably deposit a layer of material onto the semiconductor surface, it is essential that the surface is approximately flat. If no special steps were taken, this would definitely not be the case in modern CMOS processes, where multiple patterned metal interconnect layers are superimposed onto each other. Therefore, a *chemical-mechanical planarization* (CMP) step is included before the deposition of an extra metal layer on top of the insulat- ing SiO_2 layer. This process uses a slurry compound—a liquid carrier with a suspended abrasive component such as aluminum oxide or silica—to microscopically plane a device layer and to reduce the step heights.

2.1.2 Simplified CMOS Process Flow

The gross outline of a potential CMOS process flow is given in Figure 2.6. The process starts with the definition of the *active regions*, this is the regions where transistors will be constructed. All other areas of the die will be covered with a thick layer of silcion dioxide (SiO_2), called the *field oxide*. This oxide acts as the insulator between neighboring devices, and is either grown (as in the process of Figure 2.1), or deposited in etched trenches (Figure 2.2) — hence the name *trench insulation*. Further insulation is provided by the addition of a reverse-biased *np*-diode, formed by adding an extra p^+ region, called the *channel-stop implant* (or *field implant*) underneat the field oxide. Next, lightly doped *p*- and *n*-wells are formed through ion implantation. To construct an NMOS transistor in a *p*-well, heavily doped *n*-type *source* and *drain* regions are implanted (or diffused) into the lightly doped *p*-type substrate. A thin layer of SiO_2, called the

gate oxide, separates the region between the source and drain, and is itself covered by conductive polycrystalline silicon (or polysilicon, for short). The conductive material forms the *gate* of the transistor. PMOS transistors are constructed in an *n*-well in a similar fashion (just reverse *n*'s and

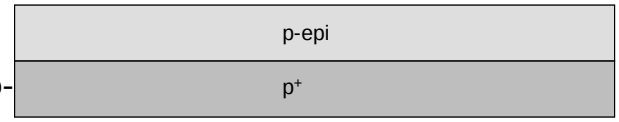


Simplified process sequence for the manufacturing of a n- dual-well CMOS circuit.

p's). Multiple insulated layers of metalic (most often Aluminum) wires are deposited on top of these devices to provide for the necessary interconnections between the transistors. A more detailed breakdown of the flow into individual process steps and their impact on the semiconductor material is shown graphically in Figure 2.7. While most of the operations should be self-explanatory in light of the previous descriptions, some com- ments on individual operations are worthwhile. The process starts with a *p*-substrate sur- faced with a lightly doped *p*-epitaxial layer (a). A thin layer of SiO_2 is deposited, which will serve as the gate oxide for the transistors, followed by a deposition of a thicker sacri- ficial silicon nitride layer (b). A plasma etching step using the complimentary of the active area mask creates the trenches, used for insulating the devices (c). After providing the channel stop implant, the trenches are filled with SiO_2 followed by a nyumber of steps to provide a flat surface (including inverse active pattern oxide etching, and chemical- mechanical planarization). At that point, the sacrificial nitride is removed (d). The *n*-well mask is used to expose only the *n*-well areas (the rest of the wafer is covered by a thich buffer material), after whihc an implant-annealing sequence is applied to adjust the well-doping. This is followed by a second implant step to adjust the threshold voltages of the PMOS transistors. This implant only impacts the doping in the area just below the gate oxide (e). Similar operations (using other dopants) are performed to created the p-wells, and to adjust the thresholds of the NMOS transistors (f). A thin layer of polysilicon is chemically deposited, and patterned with the aid of the polysilicon mask. Polysilicon is used both as gate electrode material for the transistors as well as an interconnect medium (g). Consecutive ion implantations are used to dope the source and drain

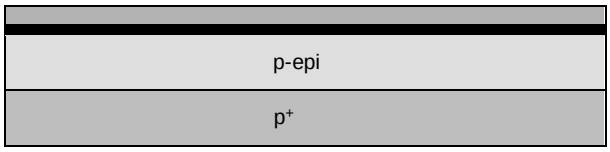
regions of the PMOS (p^+) and NMOS (n^+) transistors, respectively (h), after which the thin gate oxide not covered by the polysilicon is etched away[1]. The same implants are also use to dope the

(a) substrate with p-

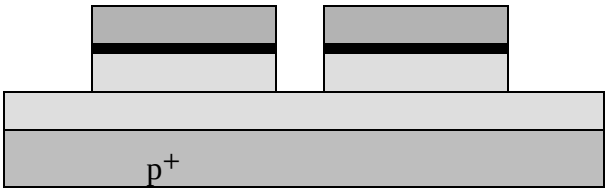


Base material: p+

Si_3N_4 SiO_2

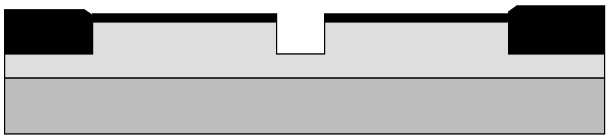


(b) After deposition of gate-oxide sacrificial nitride (acts as a buffer layer)

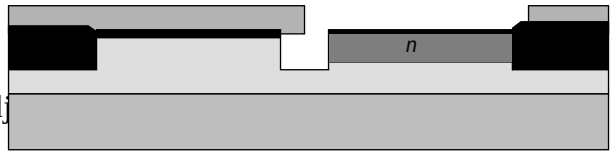


(c) After plasma etch of insulating trenches using the inverse of the active area mask

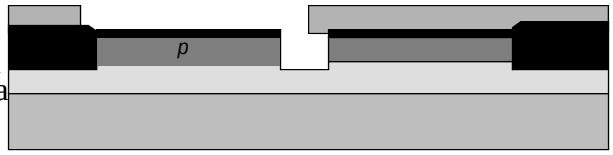
SiO_2



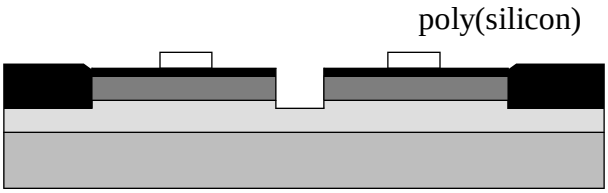
(d) After trench filling, CMP planarization, and removal of sacrificial nitride

(e) 

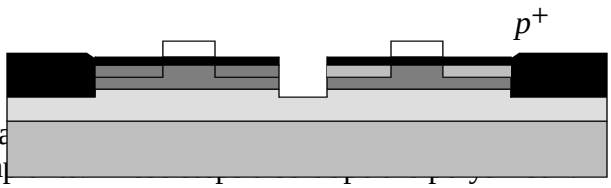
After *n*-well and V_{Tp} adj



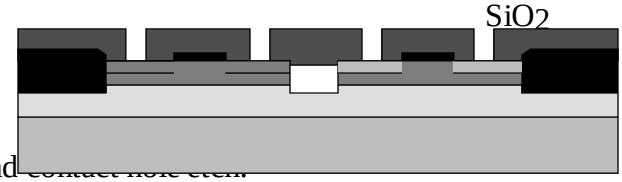
After *p*-well and V_{Tn} adjust impla



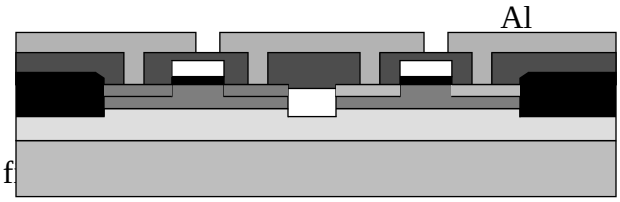
(g) After polysilicon deposition and etch



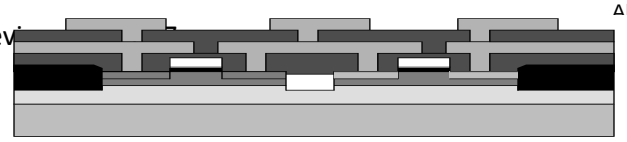
(h) After n^+ source/dra
p^+ source/drain imp



(i) After deposition of SiO_2 insulator and



(j) After deposition and patterning of f



SiO_2

(k) After deposition of SiO_2 insulator, etching of via's, deposition and patterning of second layer of Al.

Process flow for the fabrication of an NMOS and a PMOS transistor in a dual-well CMOS process. Be aware that the drawings are stylized for understanding, and that the aspects ratios are not proportioned to reality.

polysilion on the surface, reducing its resistivity. Undoped polysilicon has a very high resistivity. Note that the polysilicon gate , which is patterned before the doping, actually defines the precise location of the channel region, and hence the location of the source and drain regions. This procedure allows for a very precise positioning of the two regions rela- tive to the gate, and hence is called the *self-aligned process*. The process continues with the deposition of the metalic interconnect layers. These consists of a repition of the fol- lowing steps (i-k): deposition of the insulating material (most often $Si0_2$), etching of the contact or via holes, deposition of the metal (most often Aluminum, although Tungsten is often used for the lower layers), and patterning of the metal. Intermediate planarization
steps ensure that the surface remains reasonable flat, even in the presence of multiple interconnect layers. After the last level of metal is deposited, a final passivation or *over- glass* is deposited for protection. The layer would be CVD SiO_2, although often an addi- tional layer of nitride is deposited as it is more impervious to moisture. The final processing step is to etch openings to the pads used for bonding.

A cross-section of the final artifact is shown in Figure 2.8. Observe how the transis- tors occupy only a small fraction of the total height of the structure. The interconnect lay- ers take up the majority of the vertical dimension.

Design Rules — The Contract between Designer and Process Engineer

As processes become more complex, requiring the designer to understand the intricacies of the fabrication process and interpret the relations between the different masks is a sure road to trouble. The goal of defining a set of design rules is to allow for a ready translation of a circuit concept into an actual geometry in silicon. The design rules act as the interface or even the contract between the circuit designer and the process engineer.

Circuit designers in general want tighter, smaller designs, which lead to higher per- formance and higher circuit density. The process engineer, on the other hand, wants a reproducible and high-yield process. Design rules are, consequently, a compromise that attempts to satisfy both sides.

The design rules provide a set of guidelines for constructing the various masks needed in the patterning process. They consist of minimum-width and minimum-spacing constraints and requirements between objects on the same or on different layers.

The fundamental unity in the definition of a set of design rules is the *minimum line width*. It stands for the minimum mask dimension that can be safely transferred to the semiconductor material. In general, the minimum line width is set by the resolution of the patterning process, which is most commonly based on optical lithography. More advanced approaches use electron-beam or X-ray sources that offer a finer resolution, but are less attractive from an economical viewpoint.

Even for the same minimum dimension, design rules tend to differ from company to company, and from process to process. This makes porting an existing design between dif- ferent processes a time-consuming task. One approach to address this issue is to use advanced CAD techniques, which allow for migration between compatible processes. Another approach is to use *scalable design rules*. The latter approach, made popular by Mead and Conway [Mead80], defines all rules as a function of a single parameter, most often called □ .The rules are chosen so that a design is easily ported over a cross section of industrial processes. Scaling of the minimum dimension is accomplished by simply changing the value of □ .This results in a *linear scaling* of all dimensions. For a given pro- cess, □ is set to a specific value, and all design dimensions are consequently translated into absolute numbers. Typically, the minimum line width of a process is set to 2□ . For instance, for a 0.25 □ m process (i.e., a process with a minimum line width of 0.25 □ m), □ equals 0.125□ m.

This approach, while attractive, suffers from some disadvantages:

1. Linear scaling is only possible over a limited range of dimensions (for instance, between 0.25 □ m and 0.15 □ m). When scaling over larger ranges, the relations between the different layers tend to vary in a nonlinear way that cannot be ade- quately covered by the linear scaling rules.

2. Scalable design rules are conservative. As they represent a cross section over differ- ent technologies, they have to represent the worst-case rules for the whole set. This results in overdimensioned and less-dense designs.

For these reasons, scalable design rules are normally avoided by industry. As circuit den- sity is a prime goal in industrial designs, most semiconductor companies tend to use *micron rules*, which express the design rules in absolute dimensions and can therefore exploit the features of a given process to a maximum degree. Scaling and porting designs between technologies under these rules is more demanding and has to be performed either manually or using advanced CAD tools.

For this textbook, we have selected a "vanilla" 0.25 □ m CMOS process as our pre- ferred implementation medium. The rest of this section is devoted to a short introduction and overview of the design rules of this process, which fall in the micron-rules class. A complete design-rule set consists of the following entities: a set of layers, relations between objects on the same layer, and relations between objects on different layers. We discuss each of them in sequence.

Layer Representation

The layer concept translates the intractable set of masks currently used in CMOS into a simple set of conceptual layout levels that are easier to visualize by the circuit designer. From a designer's viewpoint, all CMOS designs are based on the following entities:

- S*ubstrates* and/or *wells*, being *p*-type (for NMOS devices) and *n*-type (for PMOS)
- *Diffusion regions* (n^+ and p^+) defining the areas where transistors can be formed. These regions are often called the *active areas*. Diffusions of an inverse type are needed to implement contacts to the wells or to the substrate. These are called *select regions*.
- One or more *polysilicon* layers, which are used to form the gate electrodes of the transistors (but serve as interconnect layers as well).
- A number of *metal interconnect* layers.
- *Contact and via* layers to provide interlayer connections.

A layout consists of a combination of polygons, each of which is attached to a certain layer. The functionality of the circuit is determined by the choice of the layers, as well as the interplay between objects on different layers. For instance, an MOS transistor is formed by the cross section of the diffusion layer and the polysilicon layer. An intercon- nection between two metal layers is formed by a cross section between the two metal lay- ers and an additional contact layer. To visualize these relations, each layer is assigned a standard color (or stipple pattern for a black-and-white representation). The different lay- ers used in our CMOS process are represented in Colorplate 1 (color insert).

Interlayer Constraints

A first set of rules defines the minimum dimensions of objects on each layer, as well as the minimum spacings between objects on the same layer. All distances are expressed in □ m. These constraints are presented in a pictorial fashion in Colorplate 2.

Interlayer Constraints

These rules tend to be more complex. The fact that multiple layers are involved makes it harder to visualize their meaning or functionality. Understanding layout requires the capa- bility of translating the two-dimensional picture of the layout drawing into the three- dimensional reality of the actual device. This takes some practice.
We present these rules in a set of separate groupings.

1 *Transistor Rules* (Colorplate 3). A transistor is formed by the overlap of the active and the polysilicon layers. From the intralayer design rules, it is already clear that the minimum length of a transistor equals 0.24 □ m (the minimum width of polysili- con), while its width is at least 0.3 □ m(the minimum width of diffusion). Extra rules include the spacing between the active area and the well boundary, the gate overlap of the active area, and the active overlap of the gate.

2 *Contact and Via Rules* (Colorplates 2 and 4). A contact (which forms an intercon- nection between metal and active or polysilicon) or a via (which connects two metal layers) is formed by overlapping the two interconnecting layers and providing a contact hole, filled with metal, between the two. In our process, the minimum size of the contact hole is 0.3 □ m, while the polysilicon and diffusion layers have to extend at least over 0.14 □ m beyond the area of the contact hole. This sets the minimum area of a contact to 0.44 □ m □ □ □ □ □ □ m□ This is larger than the dimensions of a min- imum-size transistor! Excessive changes between interconnect layers are thus to be avoided. The figure, furthermore, points out the minimum spacings between contact and via holes, as well as their relationship with the surrounding layers.

Well and Substrate Contacts (Colorplate 5). For robust digital circuit design, it is important for the well and substrate regions to be adequately connected to the supply volt- ages. Failing to do so results in a resistive path between the substrate contact of the transis- tors and the supply rails, and can lead to possibly devastating parasitic effects, such as latchup. It is therefore advisable to provide numerous substrate (well) contacts spread over the complete region. To establish an ohmic contact between a supply rail, implemented in metal1, and a *p*-type material, a p^+ diffusion region must be provided. This is enabled by the *select* layer, which reverses the type of diffusion. A number of rules regarding the use of the *select layer* are illustrated in Colorplate 5.

Consider an *n*-well process, which implements the PMOS transistors into an *n*-type well diffused in a *p*-type material. The nominal diffusion is p^+. To invert the polarity of the diffusion, an *n*-select layer is provided that helps to establish the n^+ diffusions for the well- contacts in the *n*-region as well as the n^+ source and drain regions for the NMOS transis- tors in the substrate.

Verifying the Layout

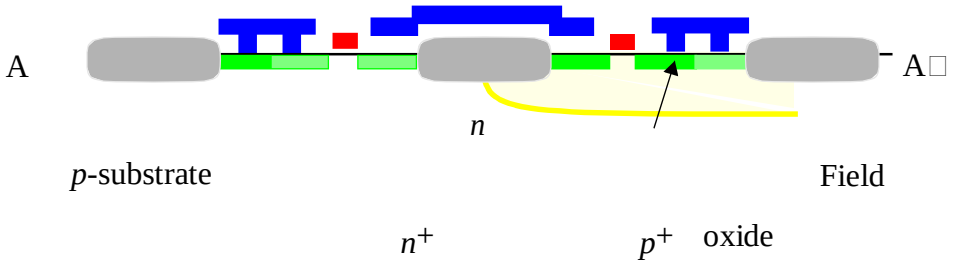
Ensuring that none of the design rules is violated is a fundamental requirement of the design process. Failing to do so will almost surely lead to a nonfunctional design. Doing so for a complex design that can contain millions of transistors is no sinecure, especially when taking into account the complexity of some design-rule sets. While design teams used to spend numerous hours staring at room-size layout plots, most of this task is now done by computers. Computer-aided *Design-Rule Checking* (called *DRC*) is an integral part of the design cycle for virtually every chip produced today. A number of layout tools even perform *on-line DRC* and check the design in the background during the time of con- ception.

Layout Example

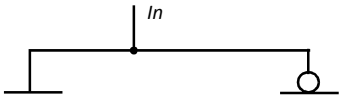
An example of a complete layout containing an inverter is shown in Figure 2.9. To help the visualization process, a vertical cross section of the process along the design center is included as well as a circuit schematic.

It is left as an exercise for the reader to determine the sizes of both the NMOS and the PMOS transistor.

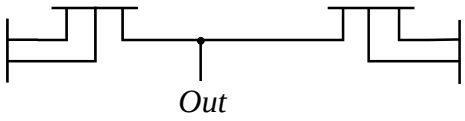
(c) Layout



(c) Cross section along A-A



GND V_{DD}



(b) Circuit diagram

Figure A detailed layout example, including vertical process cross section and circuit diagram.

Packaging Integrated Circuits

The IC package plays a fundamental role in the operation and performance of a compo- nent. Besides providing a means of bringing signal and supply wires in and out of the sili- con die, it also removes the heat generated by the circuit and provides mechanical support. Finally, its also protects the die against environmental conditions such as humidity.

The packaging technology furthermore has a major impact on the performance and power-dissipation of a microprocessor or signal processor. This influence is getting more pronounced as time progresses by the reduction in internal signal delays and on-chip capacitance as a result of technology scaling. Up to 50% of the delay of a high-perfor- mance computer is currently due to packaging delays, and this number is expected to rise. The search for higher-performance packages with fewer inductive or capacitive parasitics has accelerated in recent years.

Packaging Integrated Circuits

The increasing complexity of what can be integrated on a single die also translates into a need for ever more input-output pins, as the number of connections going off-chip tends to be roughly proportional to the complexity of the circuitry on the chip. This relationship was first observed by E. Rent of IBM (published in [Landman71]), who translated it into an empirical formula that is appropriately called *Rent's rule*. This formula relates the num- ber of input/output pins to the complexity of the circuit, as measured by the number of gates.

$$P = K \square G^{\square} \qquad (2.1)$$

where K is the average number of I/Os per gate, G the number of gates, $\square$ the Rent expo- nent, and P the number of I/O pins to the chip. $\square$ varies between 0.1 and 0.7. Its value dependsstrongly upon the application area, architecture, and organization of the circuit, as demonstrated in Table 2.1. Clearly, microprocessors display a very different input/output behavior compared to memories.

Rent's constant for various classes of systems ([Bakoglu90])

Application	β	*K*
Static memory	0.12	6
Microprocessor	0.45	0.82
Gate array	0.5	1.9
High-speed computer (chip)	0.63	1.4
High-speed computer (board)	0.25	82

The observed rate of pin-count increase for integrated circuits varies between 8% to 11% per year, and it has been projected that packages with more than 2000 pins will be required by the year 2010. For all these reasons, traditional dual-in-line, through-hole mounted packages have been replaced by other approaches such as surface-mount, ball- grid array, and multichip module techniques. It is useful for the circuit designer to be aware of the available options, and their pros and cons.

Due to its multifunctionality, a good package must comply with a large variety of requirements.

- **Electrical requirements**—Pins should exhibit low capacitance (both interwire and to the substrate), resistance, and inductance. A large characteristic impedance should be tuned to optimize transmission line behavior. Observe that intrinsic inte- grated-circuit impedances are high.
- **Mechanical and thermal properties**—The heat-removal rate should be as high as possible. Mechanical reliability requires a good matching between the thermal prop- erties of the die and the chip carrier. Long-term reliability requires a strong connec- tion from die to package as well as from package to board.
- **Low Cost**—Cost is always one of the more important properties. While ceramics have a superior performance over plastic packages, they are also substantially more expensive. Increasing the heat removal capacity of a package also tends to raise the

package cost. For instance, chips dissipating over 50 W require special heat sink attachments. Even more extreme techniques such as fans and blowers, liquid cool- ing hardware, or heat pipes, are needed for higher dissipation levels.
Packing density is a major factor in reducing board cost. The increasing pin count either requires an increase in the package size or a reduction in the pitch between the pins. Both have a profound effect on the packaging economics.

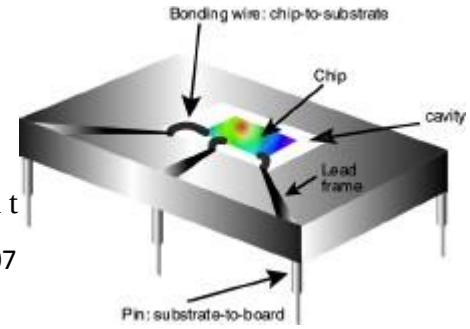
Packages can be classified in many different ways —by their main material, the number of interconnection levels, and the means used to remove heat. In this short section, we can only glance briefly at each of those issues.

Package Materials

The most common materials used for the package body are ceramic and polymers (plas- tics). The latter have the advantage of being substantially cheaper, but suffer from inferior thermal properties. For instance, the ceramic Al_2O_3 (Alumina) conducts heat better than SiO_2 and the Polyimide plastic, by factors of 30 and 100 respectively. Furthermore, its thermal expansion coefficient is substantially closer to the typical interconnect metals. The disadvantage of alumina and other ceramics is their high dielectric constant, which results in large interconnect capacitances.

2.1.3 Interconnect Levels

The traditional packaging approach uses a two-level interconnection strategy. The die is first attached to an individual chip carrier or substrate. The package body contains an internal cavity where the chip is mounted. These cavities provide ample room for many connections to the chip leads (or pins). The leads compose the second interconnect level and connect the chip to the global interconnect medium, which is normally a PC board. Complex systems contain even more interconnect levels, since boards are connected together using backplanes or ribbon cables. The first two layers of the interconnect hierar- chy are illustrated in the drawing of Figure 2.10.



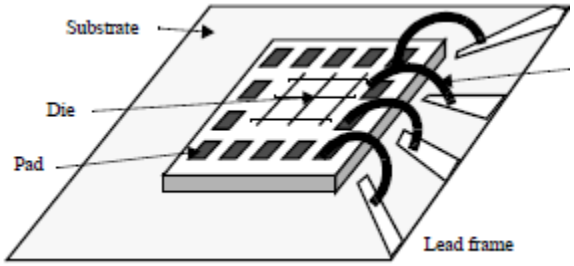
F.10 Interconnect hierarchy in t

This deep hierarchy of interconnect levels is becoming unacceptable in today's complex designs with their higher levels of integration, large signal counts, and increased performance requirements. The trend is toward reducing the number of levels. In the

future, improved manufacturing, design, and testing capabilities will make it possible to integrate a complex computer system with all its peripherals on a single piece of semicon- ductor. For the time being, attention is focused on the elimination of the first level in the packaging hierarchy. Instead of housing dies in individual packages, they are mounted directly on the interconnect medium or board. This packaging approach is called the mul- tichip module technique (or MCM), and results in a substantial increase in packing density as well as improved performance. The following sections provide a brief overview of the interconnect techniques used at levels one and two of the interconnect hierarchy, followed by a short discussion of the MCM technology.

Interconnect Level 1 —Die-to-Package-Substrate

For a long time, *wire bonding* was the technique of choice to provide an electrical connec- tion between die and package. In this approach, the backside of the die is attached to the substrate using glue with a good thermal conductance. Next, the chip pads are individually connected to the lead frame with aluminum or gold wires. The wire-bonding machine use for this purpose operates much like a sewing machine. An example of wire bonding is shown in Figure 2.11. Although the wire-bonding process is automated to a large degree, it has some major disadvantages.



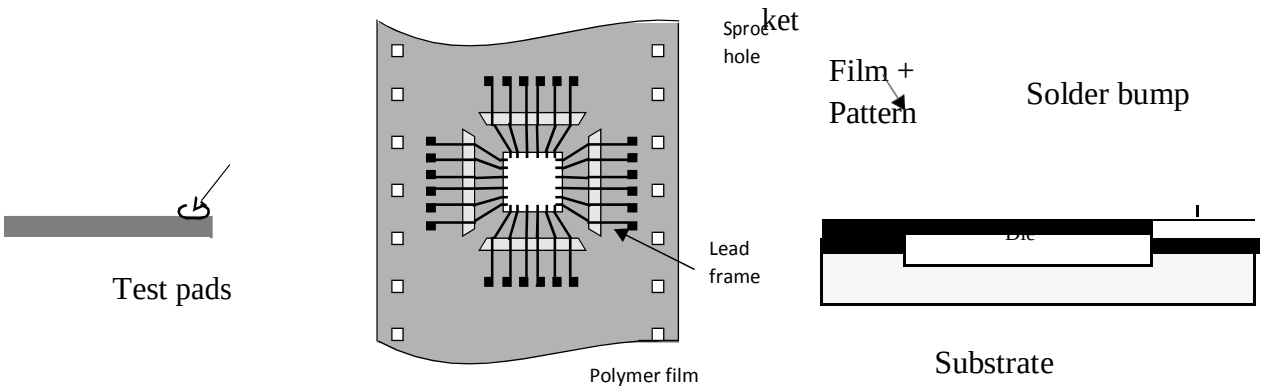
1. Wires must be attached serially, one after the other. This leads to longer manufactur- ing times with increasing pin counts.
2. Larger pin counts make it substantially more challenging to find bonding patterns that avoid shorts between the wires.
3. Bonding wires have inferior electrical properties, such as a high individual induc- tance (5 nH or

more) and mutual inductance with neighboring signals.

4. The exact value of the parasitics is hard to predict because of the manufacturing approach and irregular outlay.

New attachment techniques are being explored as a result of these deficiencies. In one approach, called *Tape Automated Bonding* (or TAB), the die is attached to a metal lead frame that is printed on a polymer film (typically polyimide) (Figure 2.12a). The con- nection between chip pads and polymer film wires is made using solder bumps (Figure 2.12b). The tape can then be connected to the package body using a number of techniques. One possible approach is to use pressure connectors.

THE MANUFACTURING PROCESS



(b) Die attachment using solder bumps

Polymer tape with imprinted wiring pattern

Figure Tape-automated bonding (TAB).

The advantage of the TAB process is that it is highly automated. The sprockets in the film are used for automatic transport. All connections are made simultaneously. The printed approach helps to reduce the wiring pitch, which results in higher lead counts. Elimination of the long bonding wires improves the electrical performance. For instance, for a two-conductor layer, 48 mm TAB Circuit, the following electrical parameters hold: *L*

Another approach is to flip the die upside-down and attach it directly to the substrate using solder bumps. This technique, called *flip-chip* mounting, has the advantage of a superior electrical performance (Figure 2.13). Instead of making all the I/O connections on the die boundary, pads can be placed at any position on the chip. This can help address the power- and clock-distribution problems, since the interconnect materials on the sub- strate (e.g., Cu or Au) are typically of a better quality than the Al on the chip.

Die

Solder bumps

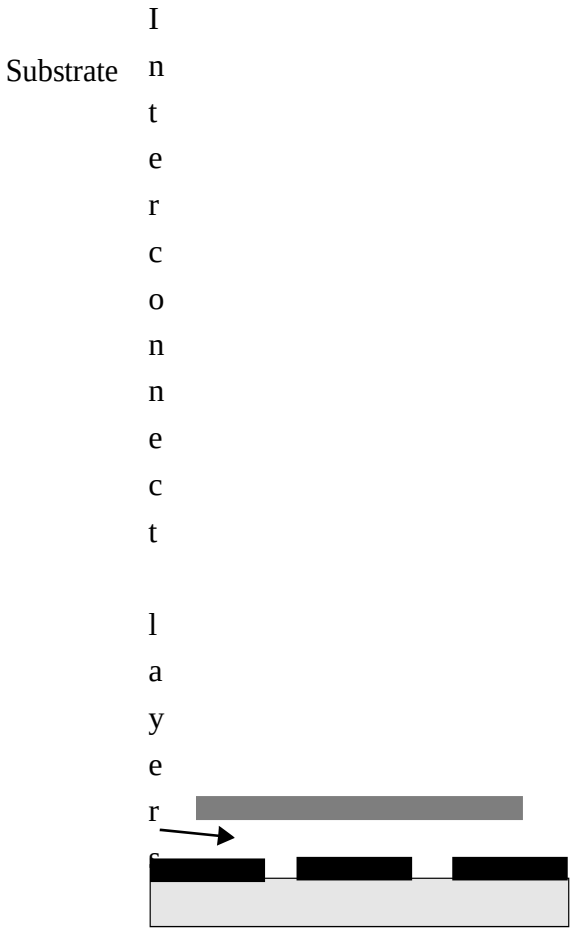
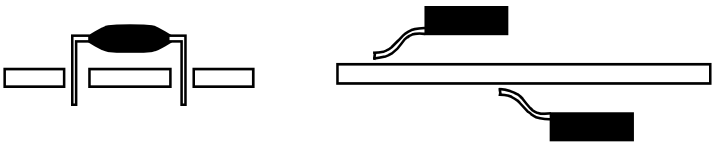


Figure 2.13 Flip-chip bonding.

Interconnect Level 2—Package Substrate to Board

When connecting the package to the PC board, *through-hole mounting* has been the pack- aging style of choice. A PC board is manufactured by stacking layers of copper and insu- lating epoxy glass. In the through-hole mounting approach, holes are drilled through the board and plated with copper. The package pins are inserted and electrical connection is made with solder (Figure 2.14a). The favored package in this class was the *dual-in-line* package or DIP (Figure 2.15a). The packaging density of the DIP degrades rapidly when the number of pins exceeds 64. This problem can be alleviated by using the *pin-grid-array*



(a) Through-hole mounting (b) Surface mount

Figure 2.14 Board-mounting approaches.

(PGA) package that has leads on the entire bottom surface instead of only on the periphery (Figure 2.15b). PGAs can extend to large pin counts (over 400 pins are possible).

The through-hole mounting approach offers a mechanically reliable and sturdy con- nection. However, this comes at the expense of packaging density. For mechanical rea- sons, a minimum pitch of 2.54 mm between the through-holes is required. Even under those circumstances, PGAs with large numbers of pins tend to substantially weaken the board. In addition, through-holes limit the board packing density by blocking lines that might otherwise have been routed below them, which results in longer interconnections. PGAs with large pin counts hence require extra routing layers to connect to the multitudes of pins. Finally, while the parasitic capacitance and inductance of the PGA are slightly lower than that of the DIP, their values are still substantial (Table 2.2).

Many of the shortcomings of the through-hole mounting are solved by using the *surface-mount* technique. A chip is attached to the surface of the board with a solder con- nection without requiring any through-holes (Figure 2.14b). Packing density is increased for the following reasons: (1) through-holes are eliminated, which provides more wiring space; (2) the lead pitch is

reduced; and (3) chips can be mounted on both sides of the board. In addition, the elimination of the through-holes improves the mechanical strength of the board. On the negative side, the on-the-surface connection makes the chip-board connection weaker. Not only is it cumbersome to mount a component on a board, but also more expensive equipment is needed, since a simple soldering iron will not do anymore□

Finally, testing of the board is more complex, because the package pins are no longer accessible at the backside of the board. Signal probing becomes hard or even impossible.

A variety of surface-mount packages are currently in use with different pitch and pin-count parameters. Three of these packages are shown in Figure 2.15: the *small-outline package* with gull wings, the *plastic leaded package* (PLCC) with J-shaped leads, and the *leadless chip carrier*. An overview of the most important parameters for a number of packages is given in Table 2.2.

Table Parameters of various types of chip carriers.

Package type	**Lead spacing (Typical)**	**Lead count (Maximum)**
Dual-in-line	2.54 mm	64
Pin grid array	2.54 mm	> 300
Small-outline IC	1.27 mm	28
Leaded chip carrier (PLCC)	1.27 mm	124
Leadless chip carrier	0.75 mm	124

Even surface-mount packaging is unable to satisfy the quest for evermore higher pin-counts. When more than 300 I/O connections are needed, solder balls replace pins as the preferred interconnect medium between package and board. An example of such a packaging approach, called ceramic *ball grid array* (BGA), is shown in . Solder bumps are used to connect both the die to the package substrate, and the package to the board. The area array interconnect of the BGA provides constant input/output density regardless of the number of total package I/O pins. A minimum pitch between solder balls of as low as
0.8 mm can be obtained, and packages with multiple 1000's of I/O signals are feasible.

Lid

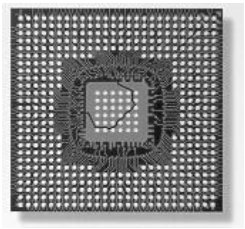
(a

Thermal grease

Chip

Flip-chip solder joints

Solder ball



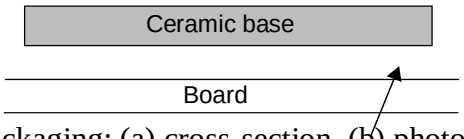


Figure Ball grid array packaging; (a) cross-section, (b) photo of package bottom

Multi-Chip Modules—Die-to-Board

Eliminating one layer in the packaging hierarchy by mounting the die directly on the wir- ing backplanes—board or substrate—offers a substantial benefit when performance or density is a major issue. A number of the previously mentioned die-mounting techniques

can be adapted to mount dies directly on the substrate. This includes wire bonding, TAB, and flip-chip, although the latter two are preferable. The substrate itself can vary over a wide range of materials, depending upon the required mechanical, electrical, thermal, and economical requirements. Materials of choice are epoxy substrates (similar to PC boards), metal, ceramics, and silicon. Silicon has the advantage of presenting a perfect match in mechanical and thermal properties with respect to the die material.

The main advantages of the MCM approach are the increased packaging density and performance. An example of an MCM module implemented using a silicon substrate (commonly dubbed *silicon-on-silicon*) is shown in Figure 2.17. The module, which imple- ments an avionics processor module and is fabricated by Rockwell International, contains 53 ICs and 40 discrete devices on a 2.2□ □ 2.2□ substrate with aluminum polyimide interconnect. The interconnect wires are only an order of magnitude wider than what is typical for on-chip wires, since similar patterning approaches are used. The module itself has 180 I/O pins. Performance is improved by the elimination of the chip-carrier layer with its assorted parasitics, and through a reduction of the global wiring lengths on the die, a result of the increased packaging density. For instance, a solder bump has an assorted capaci- tance and inductance of only 0.1 pF and 0.01 nH respectively. The MCM technology can also reduce power consumption significantly, since large output drivers—and associated dissipation—become superfluous due to the reduced load capacitance of the output pads. The dynamic power associated with the switching of the large load capacitances is simul- taneously reduced.

While MCM technology offers some clear benefits, its main disadvantage is eco- nomic. This technology requires some advanced manufacturing steps that make the pro- cess expensive. The approach is only justifiable when either dense housing or extreme performance is essential. In the near future, this argument might become obsolete as

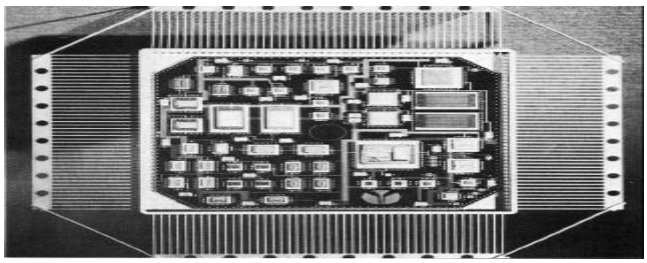


Figure Avionics processor module. *Courtesy of Rockwell International*.

MCM approaches proliferate; for example, some of the more advanced microprocessors, such as the Intel P6 (Pentium Pro), employ MCM technology.

2.1.4 Thermal Considerations in Packaging

As the power consumption of integrated circuits rises, it becomes increasingly important to efficiently remove the heat generated by the chips. A large number of failure mecha- nisms in ICs are accentuated by increased temperatures. Examples are leakage in reverse- biased diodes, electromigration, and hot-electron trapping. To prevent failure, the temper- ature of the die must be kept within certain ranges. The supported temperature range for commercial devices during operation equals 0□ to 70□ C. Military parts are more demand ing and require a temperature range varying from –55□ to 125□ C.

The cooling effectiveness of a package depends upon the thermal conduction of the package material, which consists of the package substrate and body, the package composi- tion, and the effectiveness of the heat transfer between package and cooling medium. Standard packaging approaches use still or circulating air as the cooling medium. The transfer efficiency can be improved by adding finned metal heat sinks to the package. More expensive packaging approaches, such as those used in mainframes or supercomput- ers, force air, liquids, or inert gases through tiny ducts in the package to achieve even greater cooling efficiencies.

As an example, a 40-pin DIP has a thermal resistance of 38 □ C/W and 25 □ C/W for natural and forced convection of air. This means that a DIP can dissipate 2 watts (3 watts) of power with natural (forced) air convection, and still keep the temperature difference between the die and the environment below 75 □ C. For comparison, the thermalresistance of a ceramic PGA ranges from 15 □ to 30□ C/W.

Since packaging approaches with decreased thermal resistance are prohibitively expensive, keeping the power dissipation of an integrated circuit within bounds is an eco- nomic necessity. The increasing integration levels and circuit performance make this task nontrivial. An interesting relationship in this context has been derived by Nagata [Nagata92]. It provides a bound on the integration complexity and performance as a func- tion of the thermal parameters where N_G is the

number of gates on the chip, t_p the propagation delay, ΔT the maximum temperature difference between chip and environment, θ the thermal resistance between them, and E the switching energy of each gate.

Perspective — Trends in Process Technology

Modern CMOS processes pretty much track the flow described in the previous sections although a number of the steps might be reversed, a single well approach might be fol- lowed, a grown field oxide instead of the trench approach might be used, or extra steps such as LDD (Lightly Doped Drain) might be introduced. Also, it is quite common to cover the polysilicon interconnections as well as the drain and source regions with a *sili- cide* such as $TiSi_2$ to improve the conductivity (see Figure 2.2). This extra operation is inserted between steps *i* and *j* of our process. Some important modifications or improve- ments to the technology are currently under way or are on the horizon, and deserve some attention. Beyond these, it is our belief that no dramatic changes, breaking away from the described CMOS technology, must be expected in the next decade.

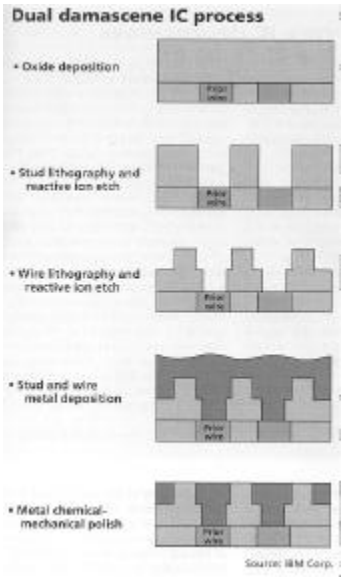
Short-Term Developments Copper and Low-k

Dielectrics

A recurring theme in this text book will be the increasing impact of interconnect on the overall design performance. Process engineers are continuously evaluating alternative options for the traditional 'Aluminum conductor—SiO_2 insulator' combination that has been the norm for the last decades. In 1998, engineers at IBM introduced an approach that finally made the use of Copper as an interconnect material in a CMOS process viable and economical [IEEESpectrum98]. Copper has the advantage of have a resistivity that is sub- stantially lower than Aluminum. Yet it has the disadvantage of easy diffusion into silicon, which degrades the characteristics of the devices. Coating the copper with a buffer mate- rial such as Titanium Nitride, preventing the diffusion, addresses this problem, but requires a special depositon process. The Dual Damascene process, introduced by IBM, (Figure 2.18) uses a metallization approach that fills trenches etched into the insulator, fol- lowed by a chemical-mechnical polishing step. This is in contrast with the the traditional approach that first deposits a full metal layer, and removes the redundant materail through etching.

In addition to the lower resistivity interconnections, insulator materials with a lower dielectric constant than SiO_2 —and hence lower capacitance— have also found their way into the production process starting with the 0.18 □ m CMOS process generation,.

(b)



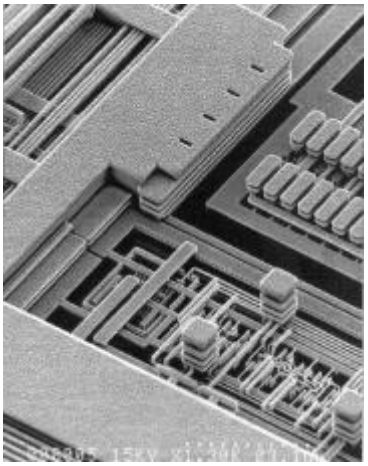


Figure 2.18

The damascene process (from IEEESpectrum98]):

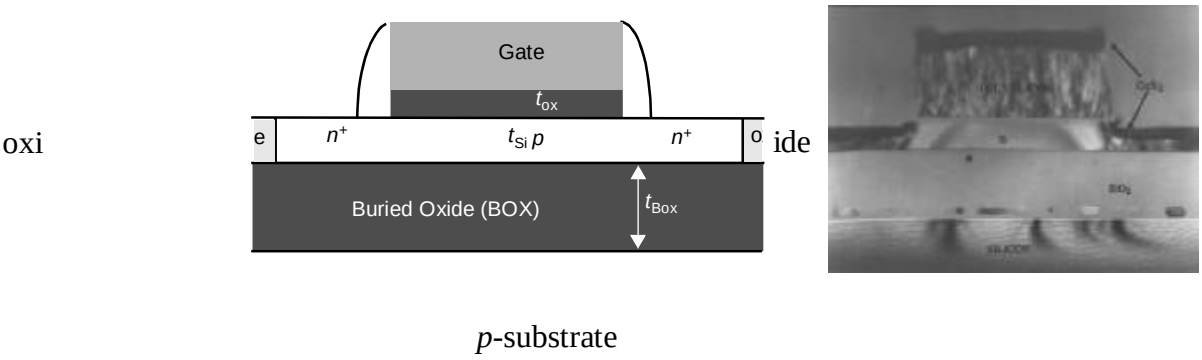
process steps (a), and microphotograph of interconnect after removal of insulator (b)

Silicon-on-Insulator

While having been around for quite a long time, there seems to be a good chance that Sili- con-on-Insulator (SOI) CMOS might replace the traditional CMOS process, described in the previous sections (also known as the *bulk CMOS process*). The main difference lies in the start material: the transistors are constructed in a very thin layer of silicon, deposited on top of a thick layer of insulating SiO_2 (Figure 2.19). The primary advantages of the SOI process are reduced parasitics and better on-off characteristics. It has, for instance, been demonstrated by researchers at IBM that the simple porting of a design from a bulk CMOS to an SOI process —leaving all other design and process parameters such as chan- nel length and oxide thickness identical— yields a performance improvement of 22% [Allen99]. Preparing a high quality SOI substrate at an economical cost was long the main hindrance against a large-scale introduction of the process. This picture has changed at the end of the nineties, and SOI is steadily moving into the mainstream.

2.1.5 In the Longer Term

Extending the life of CMOS technology beyond the next decade, and deeply below the 100 nm channel length region however will require re-engineering of both the process technology and the device structure. While projecting what approaches will dominate in that era equals resorting to crystal-ball gazing, some interesting developments are worth mentioning.



(a) (b)

Figure 2.19 Silicon-on-insulator process— schematic diagram (a) and SEM cross-section (b).

Vertical Transistors

Even while the addition of of many metal layers has turned the integrated circuit into a truly three-dimensional artifact, the transistor itself is still mostly laid out in a horizontal plane. This forces the device designer to jointly optimize packing density and performance parameters. By rotating the device so that the drain ends up on top, and the source at the bottom, these concerns are separated: packing density still is dominated by horizontal dimensions, while performance issues are mostly determined by vertical spacings (Figure 2.20). Operational devices of this type have been fabricated with channel lenghts substan- tially below 0.1 □ m.[LucentRef].

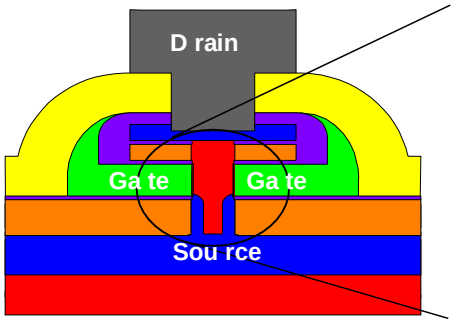


Figure Vertical transistor with dual gates. The photo on the right shows an enlarged view of the channel area.

Summary

This chapter has presented an a birds-eye view on issues regarding the manufacturing and packaging of CMOS integrated circuits.

- The manufacturing process of integrated circuits require a large number of steps, each of which consists of a sequence of basic operations. A number of these steps and/or operations, such as photolithograpical exposure and development, material deposition, and etching, are executed very repetitively in the course of the manufac- turing process.