Semiconductor Devices

Dhruva Hegde

1 Semiconductor Fundamentals

Semiconductors are the root behind all electronic devices. Before understanding the working of various devices, it is important to get a grasp on some of the concepts regarding semiconductor physics.

1.1 Density of States

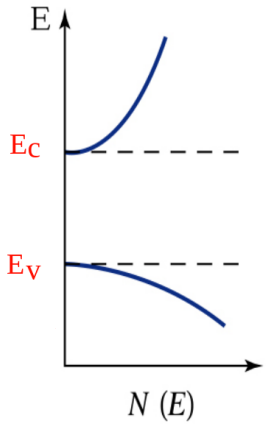
Density of states is the number of energy levels per unit energy per unit volume of a semiconductor material and is denoted as $N(E)$.
These will be solutions to the Schrodinger equation.

For 3-dimensional semiconductor crystals, the expression for density of states is found to be

$$N(E) = \begin{cases} \frac{\sqrt{(2m^3)(E-E_c)}}{\pi^2\hbar^3} & : E_c \leq E < \infty \\ \frac{\sqrt{(2m^3)(E_v-E)}}{\pi^2\hbar^3} & : -\infty < E \leq E_v \\ 0 & : E_v < E < E_c \end{cases}$$

Hence, the density of states is proportional to the square root of energy.

It is found that for 2-D crystals, DoS has no dependency on energy. For 1-D crystals, DoS is inversely proportional to square root of energy and for 0-D crystal (just 1 point), there is only 1 energy level and hence there are exactly 2 solutions (of opposite spin).

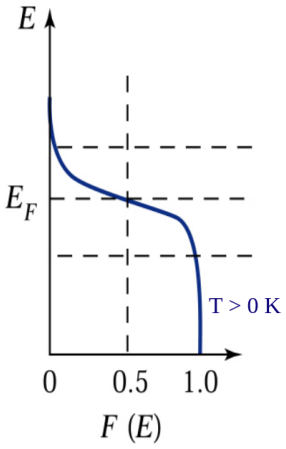


1.2 Fermi Function

Fermi Function is the probability of finding that an energy state at the given energy E is occupied by an electron. It is denoted by $f(E)$.

$$f(E) = \frac{1}{1 + e^{(E-E_f)/kT}}$$

where E_f is the Fermi level, k is the Boltzmann constant and T is absolute temperature



Fermi level is the energy level which has probability of occupancy equal to 0.5 at any temperature greater than absolute 0 (i.e $0K$).

At $T = 0K$, $f(E > E_f) = 1$ and $f(E < E_f) = 0$.

Since $f(E)$ gives the probability of occupancy of electron, $1 - f(E)$ gives the probability of non-occupancy of electron or occupancy of hole (which will be introduced soon).

Boltzmann Approximation:

- If $E - E_f >> kT$, then $f(E) \approx e^{-(E-E_f)/kT}$
- If $E - E_f << kT$, then $1 - f(E) \approx e^{(E-E_f)/kT}$

1.3 Carrier Concentration

Carriers are the charges that can contribute to current flow in a semiconductor. Electrons in the conduction band and holes in the valence band are the charge carriers.

$n(E)$ gives the number of electrons per unit volume between energy levels E and $E + dE$ (in the conduction band).
$p(E)$ gives the number of holes per unit volume between energy levels E and $E + dE$ (in the valence band).

$$\implies n(E) = N_c(E)\ f(E)\ dE \quad \text{and} \quad p(E) = N_v(E)\ (1 - f(E))\ dE$$

Total number of electrons in the conduction band is given by,

$$n = \int_{E_c}^{\infty} N_c(E)\ f(E)\ dE = N_c e^{-(E_c - E_f)/KT}$$

where N_c is the effective DoS in CB

Total number of holes in the valence band is given by,

$$p = \int_{-\infty}^{E_v} N_v(E)\ (1 - f(E))\ dE = N_v e^{-(E_f - E_v)/KT}$$

where N_v is the effective DoS in VB

1.3.1 Intrinsic Semiconductor

An intrinsic semiconductor is simply a pure semiconductor material with no additional impurities in the crystal.

In an intrinsic semiconductor, the number of electrons in the conduction band is equal to the number of holes in the valence band. $\implies n = p = n_i$ where n_i is called "Intrinsic carrier concentration".

By multiplying the expressions for n and p, the intrinsic carrier concentration is found to be

$$n_i = \sqrt{N_c\ N_v} e^{-E_g/2kT}$$

where E_g is the band gap

The bang gap is generally taken as a material constant, but it is actually a function of temperature.

$$E_g(T) = E_g(0) - \alpha\ \frac{T^2}{T + \beta}$$

where T is the absolute temperature (in K), α and β are material constants.

The Fermi level position for an intrinsic semiconductor is called "Intrinsic Fermi level" denoted by E_{fi}.

By equating the expressions for n and p, the intrinsic Fermi level is found to be

$$\boxed{E_{fi} = \frac{E_c + E_v}{2} + \frac{kT}{2} \ln\left(\frac{N_v}{N_c}\right) = \frac{E_c + E_v}{2} + \frac{3kT}{4} \ln\left(\frac{m_p^*}{m_n^*}\right)}$$

Hence, the intrinsic Fermi level lies above or below the mid point of the band gap depending on the effective masses of electrons and holes in the semiconductor material.

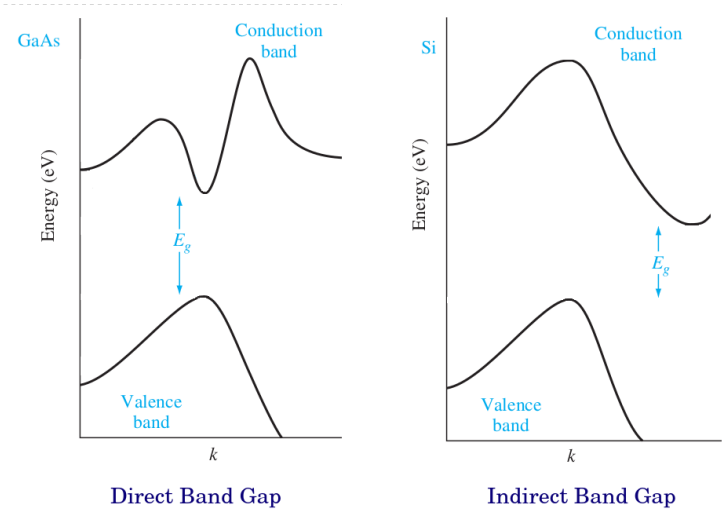
Note that effective masses of electrons and holes for Fermi level calculation are different from effective masses of electrons and holes for mobility calculation.

Types of intrinsic semiconductors:
Intrinsic semiconductors can be classified into two types based on the structure of their band gaps namely,

1. Direct bang gap semiconductors : Electrons only need to be excited with energy to make the jump from valence to conduction band and no extra momentum is necessary.

2. Indirect bang gap semiconductors : Electrons not only need to be excited with energy but they also need extra momentum in order to make the jump from valence to conduction band.

The above explanations can be observed using E vs k diagrams of the two types, which are illustrated below.



Direct Band Gap Indirect Band Gap

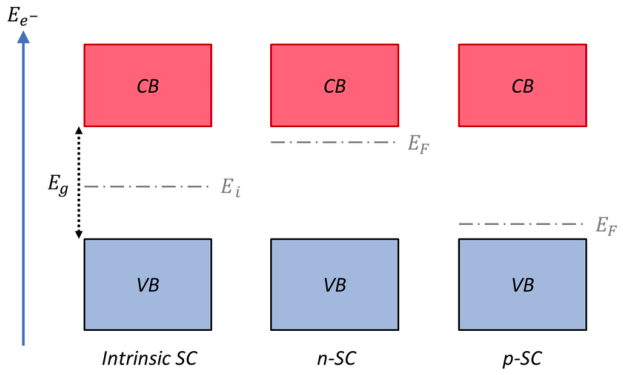
1.4 Doping

The Fermi level of an intrinsic semiconductor can be shifted by adding impurities to the crystal because it effectively changes the electron and hole concentrations.

The process of adding impurities to an intrinsic semiconductor is called "Doping" and the semiconductor obtained after doping is called "Extrinsic Semiconductor".

n-type semiconductor
Intrinsic SC is doped with penta-valent impurities i.e "donor" atoms that have 5 valence electrons.
This will increase the electron concentration and hence the Fermi level will shift above towards the conduction band.
The number of donor atoms per unit volume is represented as N_D.

p-type semiconductor
Intrinsic SC is doped with tri-valent impurities i.e "acceptor" atoms that have 3 valence electrons.
This will increase the hole concentration and hence the Fermi level will shift below towards the valence band.
The number of acceptor atoms per unit volume is represented as N_A.



Compensated doping is the case where both donor and acceptor doping is performed on an intrinsic SC.

- $N_D > N_A \implies$ n-type with $N_D - N_A$
- $N_D < N_A \implies$ p-type with $N_A - N_D$
- $N_D = N_A \implies$ intrinsic (with higher n_i)

The new Fermi levels for extrinsic semiconductors are given by,

$$E_{fn} = E_{fi} + kT \ln\left(\frac{n}{n_i}\right)$$

$$E_{fp} = E_{fi} - kT \ln\left(\frac{p}{n_i}\right)$$

Degenerate doping is the case where doping concentration is very large such that $E_c - E_f < 3kT$ or $E_f - E_v < 3kT$. In this scenario, Boltzmann approximation can't be used.

Doping will not only shift the Fermi level, but will also create disorders in the SC crystal which creates new energy states inside the band gap.
For n-type doping, new donor energy levels just below the conduction band are formed making it easier for electrons to jump into conduction band and become free.
For p-type doping, new acceptor energy levels just above the valence band are formed making it easier for electrons to jump from valence band, hence creating holes.

1.4.1 Mass Action Law

For extrinsic semiconductors, the values of n and p will not be the same.

- n-type semiconductors will have more electrons, hence electrons are the majority carriers whereas holes are the minority carriers.
- p-type semiconductors will have more holes, hence holes are the majority carriers whereas electrons are the minority carriers.

However despite this difference, at thermal equilibrium, the product of n and p will be constant and is equal to the square of the intrinsic carrier concentration. This is called "Mass Action Law".

$$\implies \boxed{n\ p = n_i^2}$$

1.4.2 Temperature Dependence

In freeze out condition ($T = 0\ K$), none of the electrons will have energy to exit valence band and move to conduction band. Hence, the number of free charge carriers at $T = 0\ K$ is zero even in extrinsic SCs.
Meaning, any SC will behave like an insulator at absolute zero.

Note that majority carriers are mostly provided by the ionized dopants while the minority carriers are provided by the intrinsic material only.

With increase in temperature, the electrons will get thermal energy to move to conduction band.

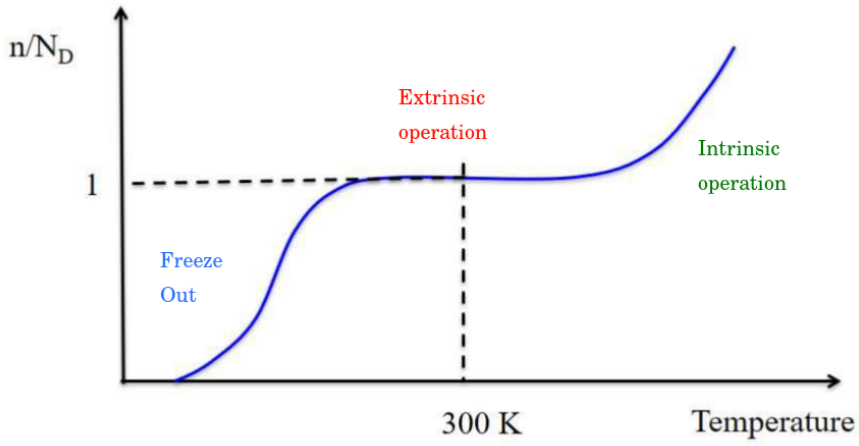
At around room temperature ($T \approx 300K$), all dopants would have ionized to give a lot of majority carriers.

As temperature increases to even higher values ($T >>> 300K$), there are no more dopants to contribute to carrier count, but contribution from intrinsic material keeps increasing till the dopant contribution becomes negligible.

- For an n-type SC
  - $T \approx 300K$: $n = N_D^+ + p \approx N_D^+$
  - $T >>> 300K$: $n \approx p$
- For a p-type SC
  - $T \approx 300K$: $p = N_A^- + n \approx N_A^-$
  - $T >>> 300K$: $p \approx n$

Hence, at around room temperature is where the doped semiconductors are said to be in "extrinsic operation region".

At higher temperatures where the dopant count is overshadowed by intrinsic carrier count, even an extrinsic SC behaves like intrinsic SC, so it is said to be in "intrinsic operation region".



The temperature at which extrinsic SC starts behaving like intrinsic SC is called "Curie Temperature".

1.4.3 Charge Neutrality Law

In general for a compensated doped extrinsic semiconductor, since the overall crystal is still electrically neutral, $p + N_D^+ = n + N_A^-$.
When all donors and acceptors are ionized, $\boxed{p + N_D = n + N_A}$.

Using the above relation and Mass Action Law, the exact values of n and p can be calculated for any extrinsic semiconductor on which compensated doping has been performed.

$$n = \frac{N_D - N_A}{2} + \sqrt{\left(\frac{N_D - N_A}{2}\right)^2 - n_i^2}$$

$$p = \frac{N_A - N_D}{2} + \sqrt{\left(\frac{N_A - N_D}{2}\right)^2 - n_i^2}$$

If there is no compensated doping and dopant concentration is much higher than intrinsic concentration, then

- n-type: $\boxed{n \approx N_D \text{ and } p \approx n_i^2/N_D}$
- p-type: $\boxed{p \approx N_A \text{ and } n \approx n_i^2/N_A}$

1.5 Generation and Recombination

Generation is the process of creation of an electron-hole pair i.e where a bounded electron in valence band is freed to conduction band.
Hence, generation increases free carrier count.

Recombination is the process of combining of a electron and a hole i.e where a free electron in conduction band drops back to valence band.
Hence, recombination decreases free carrier count.

The product of the electron and hole densities n and p is a constant i.e $np = n_i^2$ at equilibrium, maintained by recombination and generation occurring at equal rates.

When there is a surplus of carriers i.e. $np > n_i^2$, the rate of recombination becomes greater than the rate of generation, driving the system back towards equilibrium.

Likewise, when there is a deficit of carriers i.e $np < n_i^2$ the generation rate becomes greater than the recombination rate, again driving the system back towards equilibrium.

Different types of mechanisms that cause generation and recombination are

- Band to Band : direct transition from VB to CB and vice versa
- R-G centres : trap states in band gap that support jumping from VB to CB and vice versa (most effective when R-G centres/trap states are present close to the middle of the band gap)
- Excitons : e-h pairs that are bounded to each other
- Generation via impact ionization and Auger recombination

General formula for recombination rate is given by,

$$R \approx \frac{np - n_i^2}{\tau_p(n + n_i) + \tau_n(p + n_i)}$$

where τ_p and τ_n are mean life times of holes and electrons respectively

If the value of R turns out to be positive, it means recombination is dominating and if R turns out to be negative, it means generation is dominating.

Special case 1: Low level injection
Minority carriers are injected using photons (light) or something similar.

If δp is the increase in hole population in an n-type SC and δn is the increase in electron count in a p-type SC,

$$R = \begin{cases} \frac{\delta p}{\tau_p} & : n - type \\ \frac{\delta n}{\tau_n} & : p - type \end{cases}$$

Carrier concentration is above equilibrium and recombination prevails.

Note that if light is incident uniformly on a semiconductor surface at room temperature, electron-hole pairs are generated from the intrinsic material.
$\implies \delta n = \delta p$.

Special case 2: Depletion region
Depletion region is a region in which the most of the charge carriers are depleted i.e have recombined.
$\implies n << n_i$ and $p << n_i$

$$R = -\frac{n_i}{\tau_p + \tau_n}$$

Carrier concentration is below equilibrium and generation prevails.

1.5.1 Quasi Fermi levels

A semi-conductor material in equilibrium will a fixed Fermi level as illustrated earlier.
In fact, the extrinsic Fermi level of an SCs at equilibrium can be calculated using either hole or electron concentration from the formulae given and it will turn out to be the same.

However, if equilibrium is disturbed due to injection of carriers (or some other phenomenon), then the Fermi levels calculated using the concentrations will turn out to be different.

This is because, due to injection of carriers, there will not be any significant change in majority carrier concentration but there will a huge change in the minority carrier concentration, which will lead to change in value of Fermi level calculated using the minority carriers.
Such Fermi levels that are obtained due to carrier concentrations when the SC is not in equilibrium are called "Quasi Fermi levels".

1.6 Charge Transport Mechanisms

Charge transport in a semiconductor occurs in two different mechanisms which are "Drift" and "Diffusion".

Drift Current is the current caused due to electric field. The drift current density is given by,

$$J_{dr} = qnv_{dn} + qpv_{dp}$$

where q is charge of electron, v_{dn} and v_{dp} are drift velocities of electrons and holes respectively.

$$\implies \boxed{J_{dr} = q(n\mu_n + p\mu_p)\varepsilon}$$

where ε is the electric field, μ_n is mobility of electrons and μ_p is mobility of holes (since $v_p = \mu\varepsilon$).

n-type SC: $\boxed{J_{dr} = qn\mu_n\varepsilon}$
p-type SC: $\boxed{J_{dr} = qp\mu_p\varepsilon}$

Note that electron moves in direction opposite to electric field while hole moves in direction of electric field. Which is why the drift current caused due to both electrons and holes is in the same direction.

Using $J_{dr} = \sigma\varepsilon$, Conductivity of a semiconductor is given by,

$$\sigma = q(n\mu_n + p\mu_p)$$

The reciprocal of the above expression gives resistivity ρ.

Diffusion Current is the current caused due to concentration gradient along the length of the semiconductor.

Charge carriers diffuse from regions of higher concentration to regions of lower concentration.

Electron diffusion current density:

$$\boxed{J_{diff} = q\frac{dn}{dx}D_n}$$

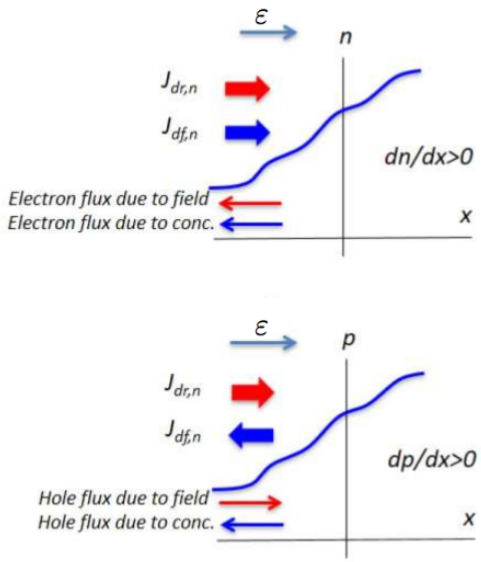
where D_n is the electron diffusion constant, which is given by $D_n = kT\mu_n/q$

Hole diffusion current density:

$$\boxed{J_{diff} = q\frac{dp}{dx}D_p}$$

where D_p is the hole diffusion constant, which is given by $D_p = kT\mu_p/q$

Note that if concentration gradient is the same, then electrons and holes diffuse in the same direction, meaning the diffusion currents caused due to electrons and holes are in opposite directions.



Total current due to both the phenomena is given by,

$$\boxed{J_n = qn\mu_n\varepsilon + qD_n\frac{dn}{dx}}$$

$$\boxed{J_p = qp\mu_p\varepsilon - qD_p\frac{dp}{dx}}$$

1.6.1 Continuity Equation

The continuity equation gives the rate of flow of charge carriers with respect to time when all mechanisms are considered.
For electrons:

$$\boxed{\frac{dn}{dt} = n\mu_n \frac{d\varepsilon}{dx} + \mu_n \varepsilon \frac{dn}{dx} + D_n \frac{d^2n}{dx^2} + G_n - R_n}$$

For holes:

$$\boxed{\frac{dp}{dt} = -p\mu_p \frac{d\varepsilon}{dx} - \mu_p \varepsilon \frac{dp}{dx} + D_p \frac{d^2p}{dx^2} + G_p - R_p}$$

(Electron flow is considered in the next few cases, analogous equations can be obtained for hole flow as well)

For low level injection of minority carriers and no electric fields: $n = n_0 + \delta n$,

$$\frac{d(\delta n)}{dt} = D_n + D_n \frac{d^2\delta n}{dx^2} + G_n - \frac{\delta n}{\tau_n}$$

Case 1: Steady state, no light
No generation due to photons and there is no change in δn with time

$$D_n \frac{d^2\delta n}{dx^2} = \frac{\delta n}{\tau_n} \qquad \implies \delta n(x) = Ae^{-x/\sqrt{D_n\tau_n}} + Be^{x/\sqrt{D_n\tau_n}}$$

where $\sqrt{D_n\tau_n} = L_n$ which is Diffusion length of electrons
(similarly $\sqrt{D_p\tau_p} = L_p$ which will be Diffusion length of holes)

Case 2: No concentration gradient, no light
No generation due to photons and there is no change in δn with space

$$\frac{d\delta n}{dt} = -\frac{\delta n}{\tau_n} \qquad \implies \delta n(t) = \delta n(0)e^{-t/\tau_n}$$

Case 3: No concentration gradient, steady state
There is no change in δn with time and space

$$G_n = \frac{\delta n}{\tau_n}$$

These special cases simplify the analysis of semiconductor behaviour under specific conditions.

2 Junctions

Before analyzing junctions, a few new terminologies need to be introduced.

Vacuum energy level (E_{vac}) is the energy level of a free stationary electron that is outside the material of interest.

Work Function (ϕ) of a material is the energy difference between the vacuum energy level and the Fermi level of the material.
Hence, it is the energy required to completely remove an electron from the material which is located at the Fermi level.

- Work function of a semiconductor is variable since Fermi level can be varied by doping
- Work function of a metal is constant

Electron Affinity (χ) is the energy difference between the vacuum energy level and the conduction band edge of a semiconductor.

- Electron Affinity of a semiconductor is a constant.

A junction is where two different types of materials are in contact on an atomic level.

When a junction is created, the Fermi levels of the materials will align due to movement of electrons (and holes). This will also result in alignment of conduction and valence bands in case of semiconductors.

Note that the alignment occurs only in the vicinity of the junction and the bulk of the materials away from the junction will be largely unaffected.

2.1 Metal-Semiconductor Junction

The work function difference between the semiconductor and the metal ($\phi_m > \phi_s$ or $\phi_m < \phi_s$) will determine the nature of a metal-semiconductor junction. Since a semiconductor can be either p-type or n-type, there are 4 different possible types of metal-semiconductor junctions.

2.1.1 Schottky Contact

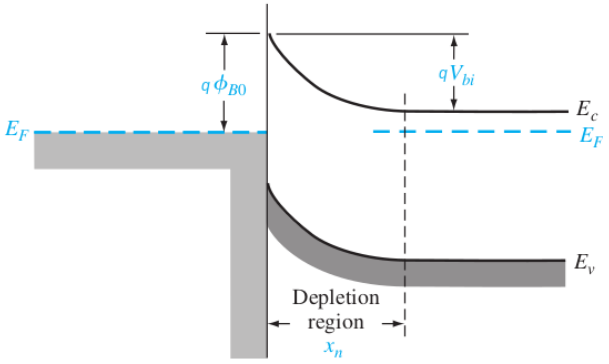
A Schottky Contact is formed from a metal-semiconductor interface in the following cases.

- the semiconductor is n-type and $\phi_m > \phi_s$
- the semiconductor is p-type and $\phi_m < \phi_s$.

Observe that in a Schottky Contact, a barrier is formed at the interface called the Schottky barrier. This will cause difficulty in flow of charge carriers through the junction.

- The barrier height ϕ_{B0} is the difference between the conduction band level edge and the Fermi level at the junction.
- The built-in voltage V_{bi} is the difference between the conduction band levels at the junction and the bulk.

Metal - n-type SC Schottky barrier:



From the band bending diagram illustrated, the following relations become apparent.

$$\phi_{B0} = \phi_m - \chi_s \qquad\qquad V_{bi} = \phi_{B0} - \phi_f$$

where $\phi_f = E_C - E_F$ at the bulk (calculated using density of states in conduction band edge and electron concentration).

The depletion region width is given by,

$$x_n = \sqrt{\frac{2\epsilon_s V_{bi}}{qN_D}}$$

The electric field developed is given by,

$$\varepsilon = \frac{qN_D(x - x_n)}{\epsilon_s}$$

The junction capacitance (per unit area) is given by,

$$C_{dep} = \frac{\epsilon_s}{x_n} = \sqrt{\frac{\epsilon_s q N_D}{2V_{bi}}}$$

Metal - p-type SC Schottky barrier:

In this case, the band bending will be such that the Schottky barrier will be pointing downwards. The behaviour will be similar, except that it applies for holes instead of free electrons.

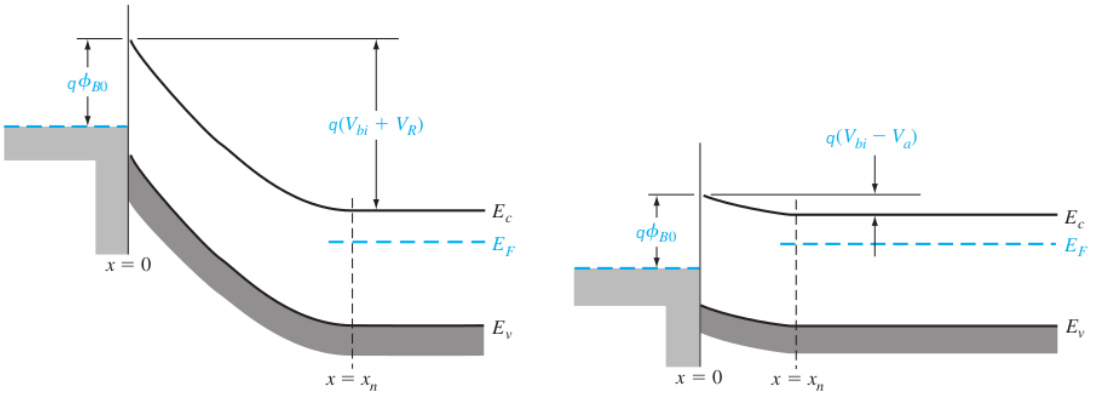
$$\phi_{B0} = \chi_s + E_g - \phi_m \qquad\qquad V_{bi} = \phi_{B0} - \phi_f$$

where E_g is the band gap and $\phi_f = E_F - E_V$ (calculated using density of states in valence band edge and hole concentration).

The other parameters are analogous, just have to replace N_D with N_A and x_n with x_p.

Biased Schottky junction: Biasing a junction corresponds to applying external voltage across it. The diagrams illustrated show how the band bending varies with applied bias voltage in a metal - n-type SC Schottky junction.

Forward bias : Metal is kept at higher potential than n-type SC (or metal is kept at lower potential than p-type SC).
Reverse bias : Metal is kept at lower potential than n-type SC (or metal is kept at higher potential than p-type SC).



The depletion width will change with biasing (which will in turn affect the junction capacitance and electric field).

$$x_n = \sqrt{\frac{2\epsilon_s(V_{bi} + V_R)}{qN_D}} \quad : \quad \text{Reverse bias}$$

$$x_n = \sqrt{\frac{2\epsilon_s(V_{bi} - V_a)}{qN_D}} \quad : \quad \text{Forward bias}$$

2.1.2 Ohmic Contact

An Ohmic Contact is formed from a metal-semiconductor interface in the following cases.

- the semiconductor is n-type and $\phi_m < \phi_s$
- the semiconductor is p-type and $\phi_m > \phi_s$.

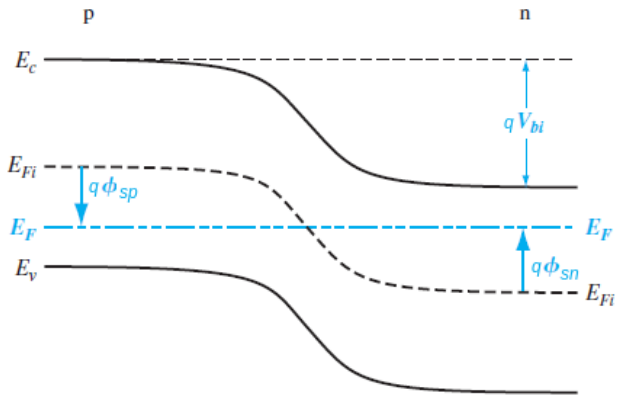
Observe that in an Ohmic Contact, there is no barrier at the interface to restrict the flow of charge carriers through the junction. In fact, flow of carriers is actually supported. Hence, an Ohmic contact can be treated like a resistor.

Due to the above reason, Ohmic contacts are preferred over Schottky contacts when metals and semiconductors have to be combined to make electronic devices. However, Schottky contacts have their own applications due to the rectifying nature.

2.2 p-n Junction

Junction formed between a p-type semiconductor and an n-type semiconductor is called "p-n Junction".

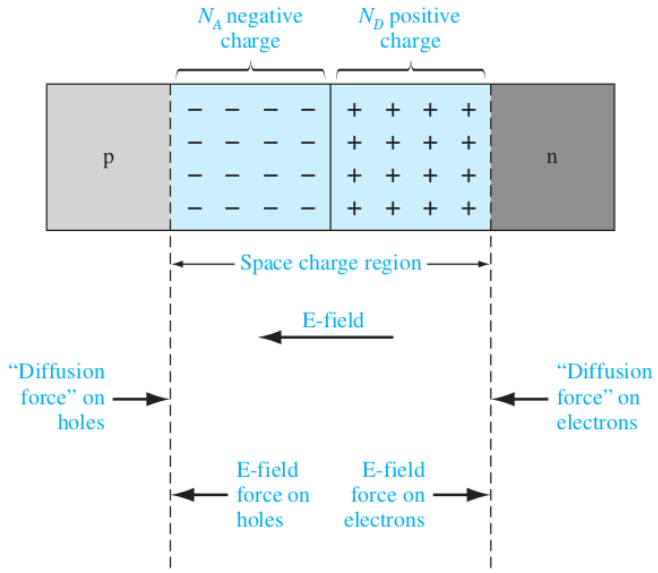
n-type : Fermi level is higher $\implies$ work function is lower
p-type : Fermi level is lower $\implies$ work function is higher
Since $\phi_{sp} > \phi_{sn}$, electrons flow from n-side to p-side and holes flow from p-side to n-side in order to align the Fermi levels.



Due to movement of free carriers from one side to another near the junction, recombination will occur and hence there will be very few free carriers left in the junction area.
This region which is depleted of charge carriers is called **Depletion Region** or "Space Charge Region".

The depletion region on p-side will have N_A^- charge per unit volume due to presence of immobile acceptor ions.
The depletion region on n-side will have N_D^+ charge per unit volume due to presence of immobile donor ions.

Due to these charges, an electric field is generated in the junction from n-side to p-side.



Electric field on the n-side:

$$\varepsilon_n = \frac{qN_D}{\epsilon_s}(x - x_{dn})$$

where x_{dn} is the width of depletion region on n-side

Electric field on the p-side:

$$\varepsilon_p = -\frac{qN_A}{\epsilon_s}(x + x_{dp})$$

where x_{dp} is the width of depletion region on p-side

At $x = 0$ i.e at the junction, $\varepsilon_n = \varepsilon_p = \varepsilon_{peak}$

$$\implies \boxed{\varepsilon_{peak} = -\frac{qN_D x_{dn}}{\epsilon_s} = -\frac{qN_A x_{dp}}{\epsilon_s}}$$

The above relation gives rise to an important result.

$$\boxed{\therefore N_D\ x_{dn} = N_A\ x_{dp}}$$

This means, the width of the depletion region on a side is inversely proportional to the doping concentration on that side.

Total width of depletion region, $x_d = x_{dn} + x_{dp}$.

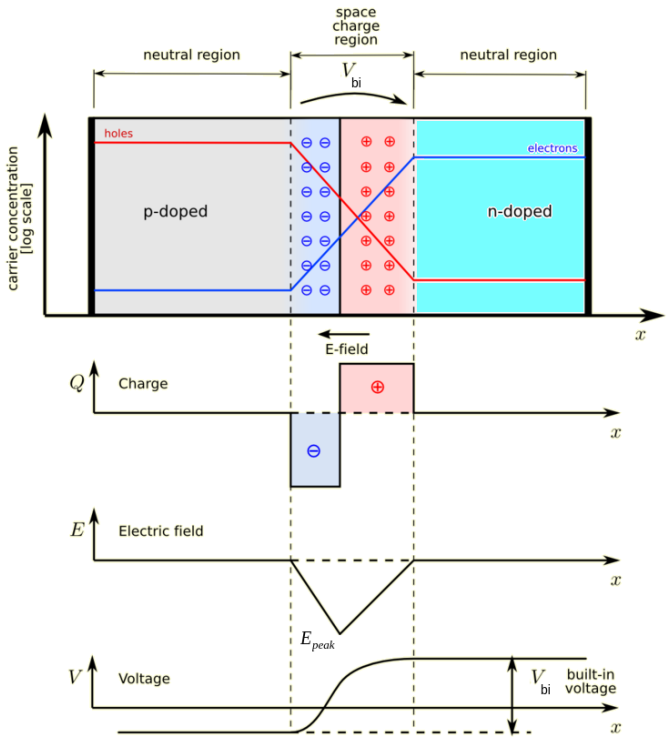
Consider Maxwell Boltzmann equation,

$$\frac{n_1}{n_2} = e^{\frac{qV_{12}}{kT}} = e^{\frac{V_{12}}{V_t}}$$

where n_1 and n_2 are carrier concentrations at two different locations and V_{12} is the potential difference between them.

If this is applied on the edges of the depletion region, the built-in potential V_{bi} developed due to band bending can be found.

$$\boxed{V_{bi} = \frac{kT}{q} \ln\left(\frac{N_A N_D}{n_i^2}\right)}$$



The width of the depletion region is found to be,

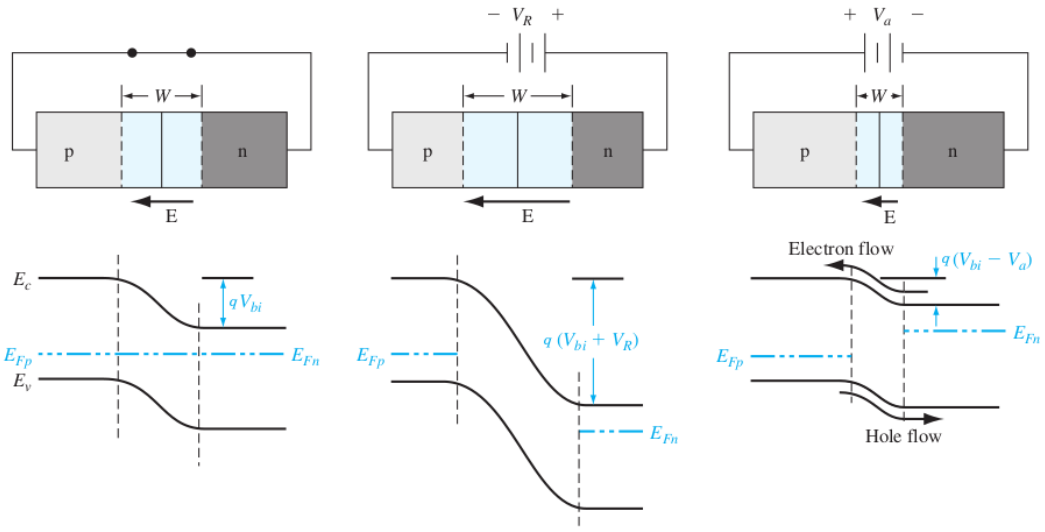
$$\boxed{x_d = \sqrt{\frac{2\epsilon_s V_{bi}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}}$$

If N_{eq} is defined as $(N_A N_D/(N_A + N_D))$, then:

$$N_D \; x_{dn} = N_A \; x_{dp} = N_{eq} \; x_d$$

2.2.1 p-n Junction Bias Conditions

Biasing is the process of applying external voltage across the p-n junction.
$V_p > V_n \implies$ Forward Bias
$V_p < V_n \implies$ Reverse Bias



Forward Biased p-n Junction
Fermi level of n-side goes up and Fermi level of p-side goes down, which causes decrease in built-in potential.

Because of this, it will be relatively easier for electrons (and holes) to diffuse from one side to another.

- Electrons diffuse from n-side to p-side

- Holes diffuse from p-side to n-side

Effectively, diffusion current flows from p-side to n-side.

Current due to drift of minority carriers is largely not affected due to forward bias voltage.
The drift current due to minority carriers will be opposite to the direction of diffusion current.

- Electrons that diffuse to p-side will create a large number of minority carriers in p-side who will eventually recombine.
- Holes that diffuse to n-side will create a large number of minority carriers in n-side who will eventually recombine.

Effectively, injection of minority carriers in both p and n regions will take place.

Continuity equation for minority carriers, low level injection under steady state, no electric field and no light can be used.

Current density at n-side of junction

$$J_p = \frac{qD_p p_{no}}{L_p} \ (e^{qV_a/kT} - 1)$$

Current density at p-side of junction

$$J_n = \frac{qD_n n_{po}}{L_n} \ (e^{qV_a/kT} - 1)$$

Hence the expression for the diffusion current density in forward biased p-n junction is given by,

$$\boxed{J = q\left[\frac{D_n n_{po}}{L_n} + \frac{D_p p_{no}}{L_p}\right] \ (e^{qV_a/kT} - 1)}$$

$qA\left[\frac{D_n n_{po}}{L_n} + \frac{D_p p_{no}}{L_p}\right]$ can be considered as constant I_0 where A is area of cross-section and thermal voltage, $V_T = kT/q$

Hence, expression for current will be

$$\boxed{I = I_0\ (e^{V_a/V_T} - 1)}$$

The above equation implies that the diffusion current through the p-n junction increases exponentially with the forward bias voltage applied.
Note that as opposed to metal-semiconductor junction, the forward current in the p-n junction is due to minority carriers.

Affect on width of depletion region:

$$x_d = \sqrt{\frac{2\epsilon_s(V_{bi} - V_a)}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}$$

Increase in forward bias voltage will decrease the depletion width.
$\implies x_d \propto \sqrt{V_{bi} - V_a}$.

Reverse Biased p-n Junction
Fermi level of n-side goes down and Fermi level of p-side goes up, which causes increase in built-in potential.

Because of this, diffusion of electrons (and holes) from one side to another will not occur as easily and hence diffusion current is negligible in case of reverse bias.

However, the increased electric field will enhance the drift of minority carriers, which is the primary cause for current in reverse biased condition.

Still, the drift current due to minority carriers in reverse bias is much less when compared to the diffusion current due to minority carriers in forward bias.

Affect on width of depletion region:

$$x_d = \sqrt{\frac{2\epsilon_s(V_{bi} + V_a)}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)}$$

Increase in reverse bias voltage will increase the depletion width.
$\implies x_d \propto \sqrt{V_{bi} + V_a}$.

2.2.2 Small signal impedance of p-n Junction

In forward bias, the impedance of a p-n junction can be modelled using-

- Series resistance offered by metal contact (r_s).
- Dynamic resistance offered by the diode to small signal variation.
  $r_d \approx \frac{kT}{qI}$
- Depletion capacitance (per unit area) that occurs due to parallel plate capacitor like structure of depletion region sandwiched between P and N sides.
  $C_{dep} = \epsilon_s / x_d$
- Diffusion capacitance (per unit area) which is caused due to fluctuation in minority carrier injection due to forward bias voltage.
  $C_{diff} = \frac{\tau I q}{kT} = \frac{\tau I}{V_T}$

Note that since $C_{dep} \propto \frac{1}{x_d}$, it will can be observed that $C_{dep} \propto \frac{1}{\sqrt{V_{bi} - V_a}}$.

In reverse bias, since currents are low series resistance will not be relevant, and nor will diffusion take place so diffusion capacitance does not exist. Hence, only depletion capacitance is prominent in reverse bias even though it is inversely proportional.

$$\boxed{C_{dep} \propto \frac{1}{(V_{bi} + V_a)^{1/n}}}$$

- $n = 2 \implies$ Step/abrupt junction
- $2 < n < 3 \implies$ Graded junction
- $n = 3 \implies$ Linearly graded junction

The above relations can be used to deduce the type of junction based on the observed change in depletion capacitance values with reverse bias voltage.

2.2.3 Non-idealities

There are other factors that will affect the performance of a p-n junction under different conditions, a few of which are discussed.

Avalanche Breakdown

At higher values of reverse bias voltage, charge carriers are forced to move at high drift velocity. These fast moving carriers can collide with atoms in the lattice and ionize them, hence creating more free charge carriers. This is called "impact ionization".
Impact ionization will in turn create more and more charge carriers, and hence high current will flow through the p-n junction. The value of reverse bias voltage at which this starts is called "Avalanche breakdown voltage".
$V_{br} \propto 1/N$ where N is doping concentration.

Ideality Factor

Due to effect of contact resistances, at higher values of applied voltage, the current increase will be slightly damped. This effect is taken care of in the diode current equation by introducing an "ideality factor" η as follows.

$$I = I_0(e^{V_a/\eta V_T} - 1)$$

$\eta = 1$ for lower values of I and $\eta = 2$ for higher values of I
There is no concrete distinction for when to use $\eta = 2$, so the general idea would be to calculate forward current using $\eta = 1$ and if it turns out to be too large to make sense in the given context, then recalculate it using $\eta = 2$.

Temperature Dependence

The reverse saturation current (or drift current) is affected by temperature because temperature will increase the number of minority carriers due to thermal excitation.

For every 10^oC rise in temperature, the current I_0 will get doubled.

$$I_0(T = T_2) = I_0(T = T_1)\ 2^{(T_2 - T_1)/10}$$

Increase in reverse saturation current will in turn increase the forward current as well.
From the above relation, it can be deduced that to maintain constant forward current, the applied voltage must reduce by around 20 mV per 10^oC increase in temperature.

Recombination & Generation

In the depletion region, when the junction is not under equilibrium, either recombination or generation will keep taking place till equilibrium condition is achieved. Hence, the overall current caused will be a combination of forward diffusion current and the current caused due to R-G mechanisms.

$$I = I_0 \ (e^{V_a/\eta V_T} - 1) \pm I_g$$

where I_g is the current caused due to generation of charge carriers
Note that if recombination is dominating, then I_r is the current caused due to recombination of charge carriers which is negative of I_g.

2.3 Special purpose Diodes

Due to their rectification property, p-n junctions diodes are very useful in electronic circuits.
Apart from that, p-n junctions can be designed in different ways and when operated in different regions, their properties can be exploited to obtain various useful applications. A few such junctions are explained in detail.

2.3.1 Zener Diode

A zener diode is a heavily doped p-n junction designed to operate in reverse bias (breakdown region).

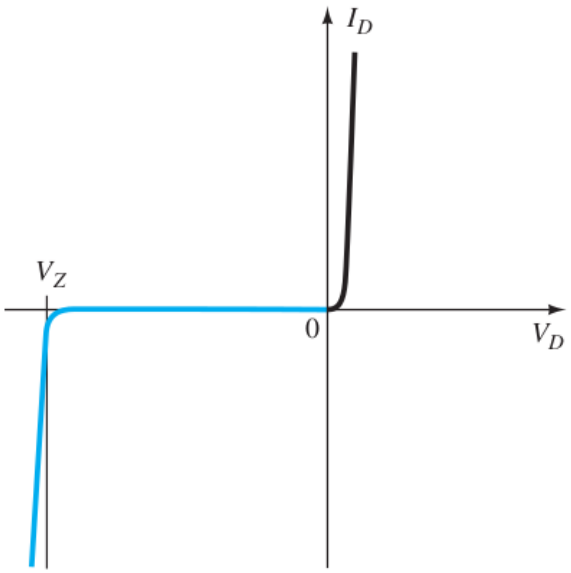
If the reverse voltage applied across the diode exceeds the breakdown voltage, the diode will act like voltage reference i.e does not allow more voltage drop across it.

The mechanism through which the zener diode enters breakdown region is different from the usual Avalanche breakdown discussed earlier.
Zener breakdown occurs due to effect of tunneling.

When both the sides are heavily doped, more charge carriers move from valence band to conduction band beyond a certain reverse voltage, which will cause the breakdown. (note that zener breakdown occurs before avalanche breakdown)

In forward bias, a zener diode acts like a normal p-n junction diode.

The knee current of a zener diode is the minimum reverse current that needs to flow through it in order for it to enter reverse breakdown region and act as voltage reference. This means the zener diode will not go to reverse breakdown region even if the reverse voltage exceeds the rating but minimum knee current is not flowing through it.



2.3.2 Solar Cell

A solar cell is a device that generates a current through some load when sunlight is incident on its surface.
Hence, solar cells work on the principle of Photo-voltaic effect.

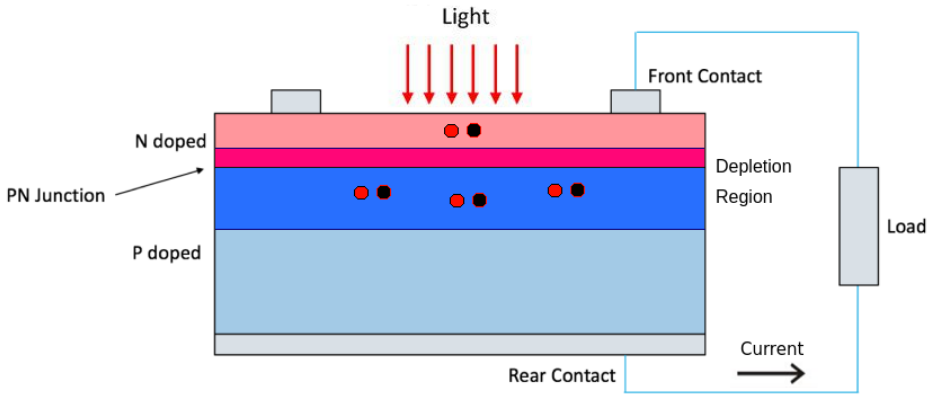
Since a solar cell is a power generating device, it does not used any bias voltage. The working mechanism of a solar cell is explained.

- Photons incident on the depletion region of the p-n junction which have more energy than the band gap will generate electron-hole pairs in the depletion region.

- Due to the electric field present in the depletion region, the generated free electrons and holes are forced to move in opposite directions, causing a current from n-side to p-side.
- Photons incident on the bulk of the material also generate e-h pairs but due to absence of electric field, they can't be driven to cause current. Hence, only photons incident on the depletion region will cause current.

Construction of solar cell:

The n-side of the p-n junction is heavily doped but thin so as to allow more photons (light) to be incident on the depletion region whereas the p-side of the p-n junction is lightly doped and large so that the depletion region width is increased.
Note that the current flowing through the load will flow from p-side to n-side since the current generated inside the cell is from n-side to p-side.

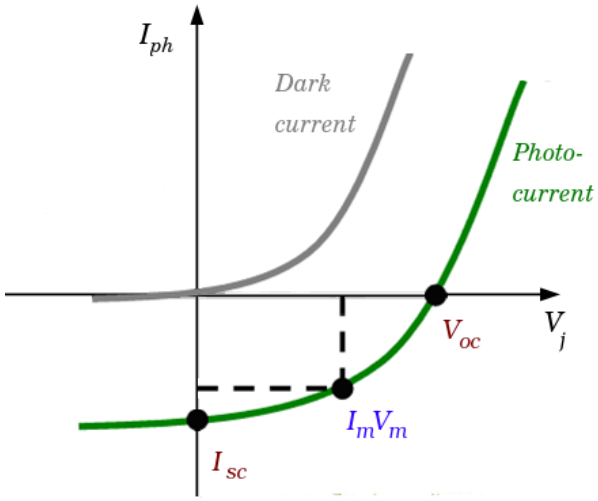


Another thing to notice is that the current flowing through the load will cause the p-n junction to self-bias itself (at some voltage V_j), which will generate some forward current.
Let the forward current be $I_f = I_0(e^{V_j/V_t} - 1)$.

The current generated due to incident photons can be expression as $I_p = qAG_nL_n$ which will be opposite to the forward current. Since electrons are the minority carriers in p-side and p-side makes up majority of the depletion region, this approximation works.

Hence the overall current is given by $\boxed{I_{ph} = I_0(e^{V_j/V_t} - 1) - qAG_nL_n}$.

This equation can be used to obtain the I-V characteristics of a solar cell.



Dark current is where there is no sunlight and the cell operates like a normal P-N junction.
When light exists, then the I-V characteristics get shifted down due to presence of photo-current in opposite direction.

1. Shorted load $\implies V_j = 0$:

   $$\text{Short circuit current} \rightarrow \quad I_{ph} = -qAG_nL_n = I_{sc}$$

2. Opened load $\implies I_{ph} = 0; \quad qAG_nL_n = I_0(e^{V_j/V_t} - 1)$

   $$\text{Open circuit voltage} \rightarrow \quad V_j = V_t \ln\left(\frac{I_{sc}}{I_o} + 1\right) = V_{oc}$$

For solar cell to be efficient, both I_{ph} and V_j should be high, but it can be observed that there is a trade-off between the two. Hence, peak efficiency is obtained when the power delivered i.e product of I_{ph} and V_j is maximum.

The maximum power that can ideally be delivered is the product of V_{oc} and I_{sc}. Practically, some operating point V_m and I_m will give maximum power.

Note that a solar cell operates in the 4^{th} quadrant i.e the developed junction voltage is positive and photo-current is negative and hence is in the opposite direction compared to forward biased p-n junction diode.

Fill Factor of a solar cell is the ratio of the power delivered at current operating point and theoretical maximum value.

$$\boxed{\text{FF} = \frac{I_m V_m}{I_{sc} V_{oc}}}$$

Efficiency of a solar cell is the ratio of power delivered to the load and the power incident on the solar cell (i.e power of incident photons).

$$\boxed{\eta = \frac{V_m I_m}{P_{opt}} = \frac{\text{FF } V_{oc} I_{sc}}{P_{opt}}}; \qquad P_{opt} = \frac{nhc}{\lambda t}$$

2.3.3 Photo-diode

A photo-diode is a device that measures the amount of radiation incident on it by generating a current proportional to the incident radiation.
Hence, photo-diode works on the principle of Photo-conductive effect.

Since a photo-diode is basically a light sensor and not a power generating device, it needs some reverse bias voltage to generate a current proportional to the incident radiation.

When photons with energies more than band gap of the semiconductor are incident on it, electron-hole pairs are generated which will cause current flow due to the bias applied.

External Quantum Efficiency of a photo-diode is the ratio of electron-hole pairs generated per unit time to the number of incident photons per unit time.
Rate of e-h pair generation is given by I_{ph}/q.
Rate of photon incidence is given by $P_{opt}\lambda/hC$.

$$\implies \eta = \frac{I_{ph} hC}{P_{opt} q \lambda}$$

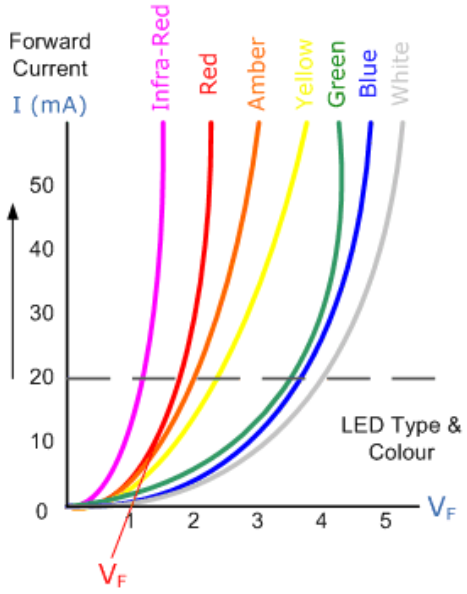
Responsivity of a photo-diode gives the current generated by the photo-diode per unit incident power of radiation.

$$R = \frac{I_{ph}}{P_{opt}} = \frac{q\lambda}{\eta hC} \qquad \Longrightarrow \boxed{\eta = \frac{1.24R}{\lambda} \quad (\lambda \text{ in } \mu m)}$$

2.3.4 Light Emitting Diode

LEDs work on the principle of electro-luminescence i.e they emit photons when forward bias voltage is applied.

LEDs are usually heavily doped p-n junction diodes made out of direct band gap semiconductors. The V-I characteristics of an LED is very similar to that of a normal p-n junction diode. LEDs allow flow of current in the forward direction i.e when forward biased and block flow of current when reverse biased, hence the region of operation will be the first quadrant.



Based on the material used to make an LED, the photons emitted will be of specific wavelengths which results in specific colors of light.

3 Transistors

Transistors are three terminal devices that are widely used as switches and for amplifiers.

A transistor can be used to work as a controlled switch. A controlled switch has 3 terminals i.e a control terminal, and two terminals that are either connected (closed) or disconnected (opened) depending on the signal applied at the control terminal.
(will be elaborated in depth in Digital Circuits)

A transistor, if biased in the correct region of operation, can also act like a small signal amplifier.
(will be elaborated in great depth in Analog Circuits)

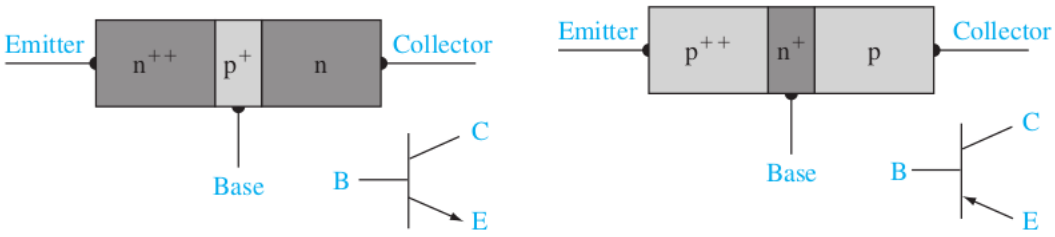
Classification of transistors:

1. **Bipolar Junction Transistor (BJT)**
2. Field Effect Transistor (FET)
   (a) Junction Field Effect Transistor (JFET)
   (b) Metal Oxide Semiconductor Field Effect Transistor (MOSFET)
      i. **Enhancement Type MOSFET (E-MOSFET)**
      ii. Depletion Type MOSFET (D-MOSFET)

3.1 Bipolar Junction Transistor (BJT)

A BJT is a three terminal, two junction device.
The terminals of a BJT are called - emitter (E), base (B) and collector (C).
The two junctions of a BJT are the emitter-base junction (EB) and collector-base junction (BC).

There are two types of BJT, which are npn BJT and pnp BJT.
npn BJT has the emitter and collector as n-type SCs with base being n-type.
pnp BJT has the emitter and collected as n-type SCs with base being n-type.

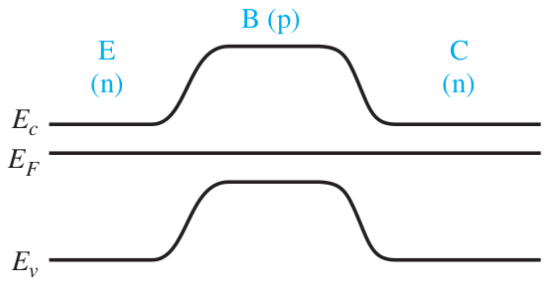


Note the following regarding size and doping concentrations:

- emitter is of moderate size and is heavily doped (so as to have more charge carriers that can create current)
- base is very thin (must be less than diffusion length of minority charge carriers in the base region) and moderately/lightly doped
- collector is wide/long (so as to give more space for heat dissipation) and is lightly/moderately doped

The npn BJT will be studied in detail since it is more widely used. The same concepts will apply to pnp BJTs as well with majority and minority charge carriers swapped.

Energy band diagram of npn BJT in equilibrium (zero bias) is shown.



3.1.1 Modes of Operation

Since a BJT has 2 junctions, based on the bias applied to each of the junction, there are 4 modes of operation.

1. EB:reverse biased; BC:reverse biased → Cut-off mode
2. EB:forward biased; BC:reverse biased → Forward active mode
3. EB:reverse biased; BC:forward biased → Inverse active mode
4. EB:forward biased; BC:forward biased → Saturation mode

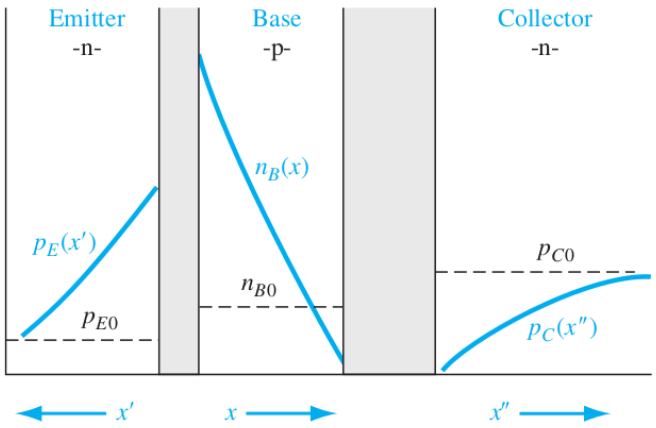
Cut-off and saturation modes are used as 'off' and 'on' states when the BJT is operated as a switch. Forward active mode (or simply active mode) is used when BJT is to be working as a small signal amplifier.

Active Mode operation:

- The forward bias on the emitter–base junction will cause current to flow across this junction. Current will consist of two components: electrons injected from the emitter into the base, and holes injected from the base into the emitter. Both these components cause emitter current I_E. Since emitter is heavily doped and base is lightly doped, the number of electrons injected is much higher than the number of holes injected.
- These electrons will be minority carriers in the base region. Because their concentration will be highest at the emitter side of the base and ideally zero at the collector side of the base which is reverse biased, there is a large concentration gradient of electrons in a very small base region.
- This is why the injected electrons will diffuse through the base region toward the collector. In their journey across the base, some of the electrons will combine with holes, which are majority carriers in the base. However, since the base is very thin i.e base width lesser than the diffusion length of electrons, the proportion of electrons that are "lost" through this recombination process will be quite small. The holes that are lost here due to recombination need to be re-substituted, which constitutes the base current I_B.
- Thus, most of the diffusing electrons will reach the boundary of the collector–base depletion region. Because the collector is more positive than the base reverse-bias voltage these successful electrons will be swept across into the collector. They will thus get collected and constitute the collector current I_C.

Every BJT will always satisfy the following equation: $I_E = I_C + I_B$.

The minority carrier distribution in an npn BJT operating in the forward-active mode is illustrated.



Calculation of current gain:
$I_{nE} \rightarrow$ Current due to the diffusion of electrons from emitter into base
$I_{pE} \rightarrow$ Current due to the diffusion of holes from base into emitter
$I_{nC} \rightarrow$ Current due to the diffusion of electrons from base into collector
$I_{RB} \rightarrow$ Difference between I_{nE} and I_{nC} , which is due to the recombination of excess electrons with carrier in the base. It represents the flow of holes into the base to replace the holes lost by recombination.

The current amplification factor α is given by the ratio of emitter current to collector current. $\alpha = \frac{I_E}{I_C}$.

$I_E = I_{nE} + I_{pE}$ and $I_C = I_{nC}$.

$$\implies \alpha = \frac{I_{nC}}{I_{nE} + I_{pE}}$$

The emitter injection efficiency (γ^*) is the ratio of emitter current that is actually contributing to the collector current to the total emitter current.

$$\gamma^* = \frac{I_{nE}}{I_{nE} + I_{pE}}$$

The base transport factor (β^*) is the ratio of collector current to the emitter current which actually contributes to the collector current.

$$\beta^* = \frac{I_{nC}}{I_{nE}}$$

From the above equations it can be deduced that,

$$\alpha = \gamma^* \ \beta^*$$

Ideally, all these values should be equal to 1.

The current gain β is given by the ratio of collector current to base current. $\beta = \frac{I_C}{I_B}$.

$$\implies \beta = \frac{\alpha}{1-\alpha} \qquad\qquad \alpha = \frac{\beta}{1+\beta}$$

Ideally, since I_C should be equal to I_E and hence $I_B = 0$, the value of β should be infinite.

Leakage currents:

- I_{CB0} $\rightarrow$ Reverse Leakage Current between Collector and Base (measured while Emitter is open).
- I_{CEO} $\rightarrow$ Reverse Leakage Current between Collector and Emitter (measured while Base is open).

The relation between them is $I_{CEO} = (\beta + 1)I_{CBO}$ (since $I_E = (\beta + 1)I_B$). These currents contribute to the actual currents as well but are usually negligibly small.

$$I_C = \alpha I_E + I_{CBO} = \beta I_B + I_{CE0} = \beta I_B + (\beta + 1)I_{CB0}$$

For design of amplifiers or other analog circuits, the general model used to analyse a BJT in active mode is described below.

- The base-emitter voltage V_{BE} has to exceed the on voltage (usually 0.7V) for the EB junction to be forward biased.

- The collector current is calculated as a function of the base-emitter voltage.

  $$I_C = I_s(e^{V_{BE}/V_t} - 1)$$

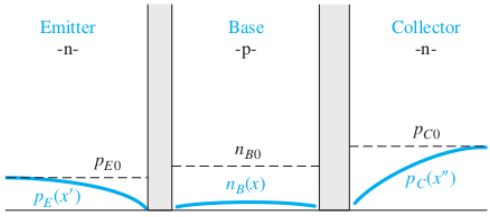
  $I_s \rightarrow$ Reverse saturation current at equilibrium

- In active region, since $I_C = \beta I_B$ and $I_E = (\beta + 1)I_B$,

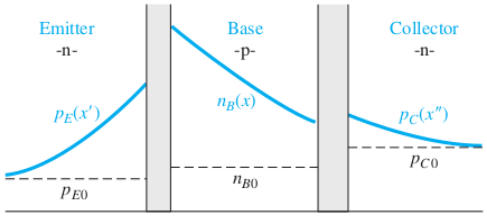
  $$I_B = \frac{I_s}{\beta}(e^{V_{BE}/V_t} - 1) \qquad I_E = \frac{I_s(\beta + 1)}{\beta}(e^{V_{BE}/V_t} - 1)$$

- It must be ensured that the CB junction is reverse biased or else the BJT will not be in active region and the above calculations will be invalid.

Cut-off Mode:
In cut-off mode, since both the junctions are reverse biased, the minority carrier concentrations are zero at both edges of depletion region.
Hence, there is not much concentration gradient of minority carriers to cause significant current.

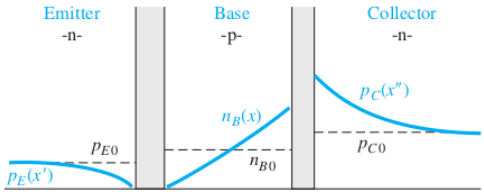


Saturation Mode:
In saturation mode, since both the junctions are forward biased, excess minority carriers exist at both edges of depletion region.



Here, maximum current flows through the transistor, as both junctions are forward biased and bulk resistance offered is very much less.
Note that in saturation region, the (approximate) linear relationship between the collector and base currents $I_C = \beta I_B$ will not hold as $I_C = I_{C_{sat}}$ which is the maximum collector current that can flow.
Hence, increase in I_B will increase I_E but I_C will remain constant.

Inverse Active Mode:
The minority carrier concentrations in inverse active mode is illustrated. It is similar to active mode, but not as efficient due to the design of the BJT (it would be exactly same as active mode if the BJT was a symmetric device i.e if collector and emitter were interchangeable).



3.1.2 Early Effect

When CB junction is reverse biased, change in collector-base voltage will change the width of the base. As noted earlier, the base width plays an important role in the operation of BJT.
With higher reverse voltage, the width of the depletion region increases which decreases the effective base width. Decrease in base width will have the following effect on BJT currents:

- Decrease in I_B due to lower recombination as size of base has decreased
- Increase in I_E due to higher diffusion as size of base has decreased
- Increase in I_C because $I_C = I_E - I_B$

Though this may seem like a good thing, it is not because Early effect gives a dependence of V_{CB} (or V_{CE}) on the collector current I_C, which is not favourable for designing amplifiers.

To model this effect in active region, the equation for collector current has to be modified.

$$I_C = I_s e^{V_{BE}/V_t}\left(1 + \frac{V_{CE}}{V_A}\right)$$

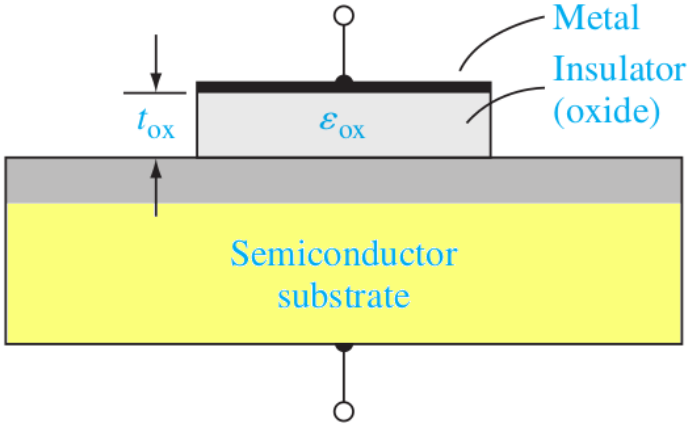
where V_A is the Early voltage

Early effect is also called "Base Width Modulation".

Punch-through is an extension of base width modulation where the reverse voltage applied causes the depletion region to widen enough to entirely occupy the base. Hence, base region won't technically exist and the BJT can't be analyzed as a two junction device.
The reverse collector-base voltage at which this occurs is called "Punch-through voltage".

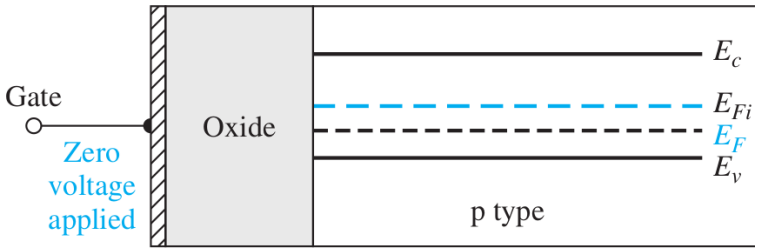
3.2 MOS Capacitor

Before understand how a metal oxide semiconductor field effect transistor works, it is important to understand the working of metal oxide semiconductor capacitor.



As shown in the figure, a MOS-cap has a metal layer and semiconductor substrate (bulk) with an insulating oxide layer sandwiched in between them. The metal layer is called "Gate" terminal.

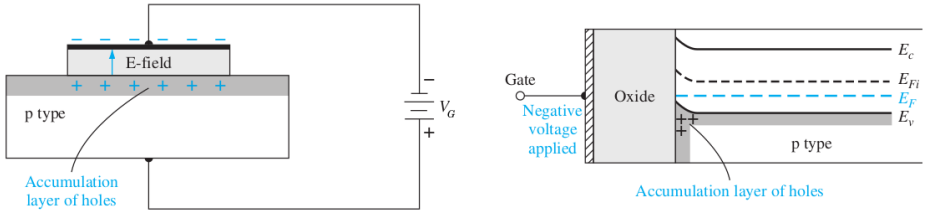
Initially for analysis purpose, it is assumed that the work function of metal and the work function of semiconductor are the same. If a p-type substrate is used, then the diagram below shows the energy levels at zero Gate voltage.



3.2.1 MOS-Cap Modes of Operation

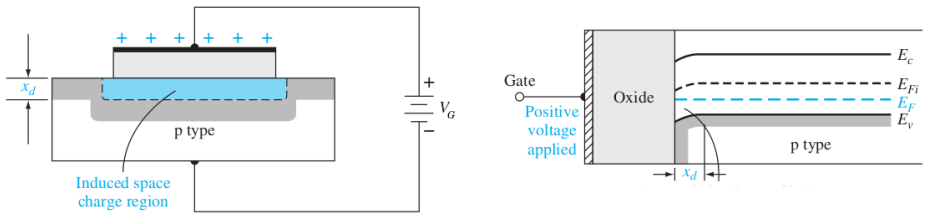
Based on the applied Gate voltage (V_G), the MOS structure operation can be divided into 3 modes.

1. **Accumulation mode**:$V_G < 0$
   When the Gate voltage is negative, then band bending causes the energy levels to shift upwards as illustrated.



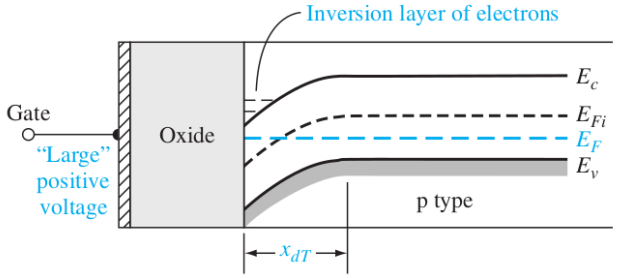
   Due to presence of negative potential on the metal side, holes will be attracted towards the oxide-semiconductor interface. This is called 'Accumulation mode' since charge carriers are accumulated at the surface.

2. **Depletion mode**: $V_G > 0$
   When the Gate voltage is positive (but less than Threshold voltage, which will be defined soon), then band bending causes energy levels to shift downwards as illustrated.



   Due to presence of positive potential on the metal side, holes will be repelled away from the oxide-semiconductor interface which causes a depletion region near the interface.
   This causes the substrate to behave like an intrinsic semiconductor near the interface.
   When the surface (built-in) potential V_{bi} becomes equal to the difference between intrinsic and p-type Fermi levels i.e $qV_{bi} = E_{Fi} - E_F$, it behaves exactly like intrinsic material.

3. **Inversion mode**: $V_G >> 0$
   When the Gate voltage is increased to higher positive values (beyond Threshold voltage), then band bending causes energy levels to shift downwards further as illustrated.



   Due to high positive potential on the metal side, minority electrons in the p-type substrate will accumulate in the depletion region near the oxide-semiconductor interface.
   This causes the substrate to behave like an n-type semiconductor near the interface, which can be observed in the band diagram.
   When the surface (built-in) potential V_{bi} becomes equal to twice the difference between intrinsic and p-type Fermi levels i.e $qV_{bi} = 2(E_{Fi} - E_F)$, it achieves inversion.
   Since a p-type substrate behaves like an n-type material near the interface, this mode is called inversion mode.

3.2.2 Threshold Voltage

The Gate voltage that needs to be applied so that the MOS capacitor goes to inversion mode of operation is called the threshold voltage.
Threshold voltage is reached when the applied Gate voltage is high enough to deplete the oxide-semiconductor interface of holes and make surface potential equal to twice the difference between intrinsic and p-type Fermi levels to achieve inversion.

$$\implies V_T = \frac{Q_{dep}}{C_{ox}} + 2\phi_f$$

$Q_{dep} = qN_A x_d \rightarrow$ Depletion charge per unit area.
$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \rightarrow$ Oxide capacitance per unit area.

Difference between intrinsic and p-type Fermi levels can be calculated as,

$$\phi_f = \frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right)$$

The depletion width x_d is given by,

$$x_d = \sqrt{\frac{2\epsilon_s V_{bi}}{qN_A}} = \sqrt{\frac{4\epsilon_s \phi_f}{qN_A}}$$

Upon substitution and re-arranging, a more fundamental expression for threshold voltage can be obtained.

$$\implies V_T = \frac{\sqrt{4q\epsilon_s \phi_f N_A}}{C_{ox}} + 2\phi_f$$

As specified earlier, this analysis was performed by assuming that the work function of metal and semiconductor are the same i.e $\phi_m = \phi_s$. However, this will not always be the case as discussed in the metal-semiconductor junction part.
Band bending will occur even before application of any Gate voltage due to this difference in work functions. Hence, to achieve inversion, the applied Gate voltage must also compensate for this band bending (could be either upwards or downwards depending on the work functions).
The voltage required to obtain flat energy bands back from bending due to work function difference is called 'flat band voltage' and this is equal to the difference between the work functions. $\implies V_{FB} = \phi_m - \phi_s$.

The final expression for threshold voltage after compensating for the work function difference is as follows.

$$\therefore V_T = V_{FB} + \frac{\sqrt{4q\epsilon_s \phi_f N_A}}{C_{ox}} + 2\phi_f$$

$$\implies V_T = \phi_m - \phi_s + \frac{\sqrt{4q\epsilon_s \phi_f N_A}}{C_{ox}} + 2\frac{kT}{q} \ln\left(\frac{N_A}{n_i}\right)$$

To get a surface potential of V_s, then the Gate voltage that needs to be applied will be,

$$V_G = \phi_m - \phi_s + \frac{\sqrt{2q\epsilon_s V_s N_A}}{C_{ox}} + V_s$$

3.2.3 Behaviour of MOS-Cap

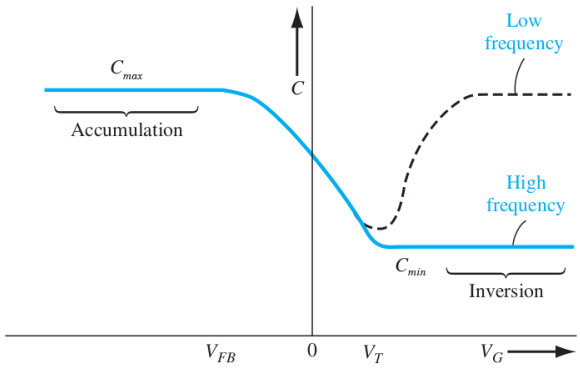
A MOS Capacitor behaves differently for different modes of operation and the behaviour also varies with frequency of operation and rate of change in Gate voltage V_G.

Low frequency AC, gradual increase in V_G

- Accumulation mode:
  Variations in Gate voltage will result in corresponding variations in hole count at the oxide-semiconductor interface.
  The effective capacitance is given by,
  $$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$$
- Depletion mode:
  Variations in Gate voltage will also result in variations in depletion width, hence there will be a depletion capacitance in series with the oxide capacitance.
  The effective capacitance is given by,
  $$\frac{C_{ox}C_{dep}}{C_{ox}+C_{dep}} = \frac{1}{\frac{t_{ox}}{\epsilon_{ox}}+\frac{x_d}{\epsilon_{ox}}}$$
- Inversion mode:
  Variations in Gate voltage will result in corresponding variations in electron count at the oxide-semiconductor interface.
  The effective capacitance is given by,
  $$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}}$$

High frequency AC, gradual increase in V_G
Behaviour is same in accumulation and depletion modes. However, in inversion mode, the electrons have to be thermally generated since they are minority carriers. The frequency of operation will be too fast for electrons to respond in time, meaning depletion region will still exist and hence the effective capacitance is same as in depletion mode.

$$C_{max} = C_{ox} \qquad C_{min} = \frac{C_{ox}C_{dep_{min}}}{C_{ox}+C_{dep_{min}}} \qquad \left[C_{dep_{min}} = \frac{\epsilon_s}{x_{d_{max}}}\right]$$



High frequency AC, sudden sweep in V_G
In this case, there is no time for inversion layer to get formed. Hence, the positive potential at V_G is only balanced by the increase in depletion width and not by inversion charge carriers. This is why the width keeps increasing which in turn causes the capacitance to keep decreasing.

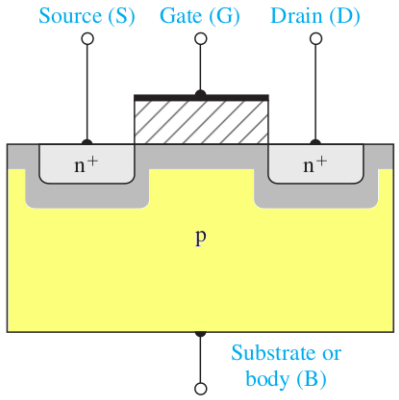
Some useful observations:

- If work function of metal is higher than work function of semiconductor, $\phi_m - \phi_s > 0$ and hence the flat band voltage V_{FB} is positive. The graph will be shifted towards the right of ideal graph (where $\phi_m = \phi_s$).
- If work function of metal is lower than work function of semiconductor, $\phi_m - \phi_s < 0$ and hence the flat band voltage V_{FB} is negative. The graph will be shifted towards the left of ideal graph (where $\phi_m = \phi_s$).
- If n-type substrate is used, p-type inversion layer will be formed and same analysis holds but all charges have to be reversed. This means the C-V curve of MOS-cap with n-type substrate will be mirror image of the one shown earlier (which was for p-type substrate), and hence the threshold voltage V_T will be negative.

3.3 MOS Field Effect Transistor (MOSFET)

As mentioned earlier, there are two types of MOSFETs, which are enhancement type MOSFET and depletion type MOSFET.
Here, the n-channel enhancement type MOSFET (nMOS) will be analysed in detail and analogies will help understand p-channel enhancement type MOSFET (pMOS) as well.
Depletion type MOSFET concepts will also be briefly covered later.

The cross-sectional view of an enhancement type nMOS is shown.



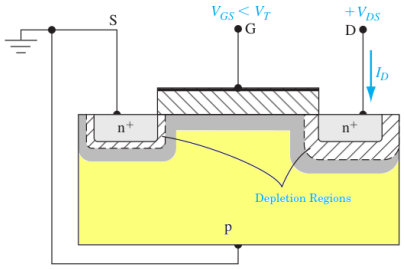
Note that the body/substrate is made of lightly doped p-type semiconductor (as in case of MOS-Cap discussed). The Gate terminal is the same. The new terminals are Drain and Source which are made out of heavily doped n-type semiconductors.
The Drain and Source terminals are interchangeable and hence a MOSFET is a symmetric device.

3.3.1 Enhancement type MOSFET: Modes of operation

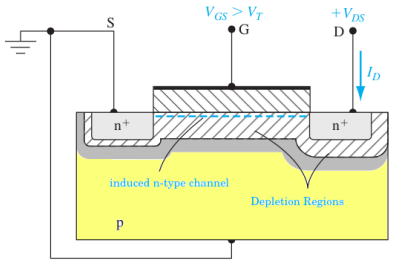
Consider n-channel enhancement type MOSFET , called nMOS in short.
The Body and Source are are internally shorted to ground. Hence, there is the applied voltage at Gate is termed V_{GS} and the applied voltage at Drain is termed V_{DS}.

The modes of operation of nMOS are relatively simpler.

Cut-off mode:
When $V_{GS} = 0$, there is no inversion channel formed between Drain and Source for current flow. In fact, the channel is formed only when V_{GS} reaches the threshold voltage V_T.
This means the nMOS is cut-off for all voltages $V_{GS} < V_T$ and there is no conduction between Drain and Source i.e $I_D = 0$.
This mode is used as 'off' state of a switch.



Linear mode:
When V_{GS} is increased beyond V_T, inversion channel is induced between Drain and Source which facilitates current flow. However, for flow of current to actually happen, a potential difference is necessary between Drain and Source.
For this purpose, V_{DS} is applied. The application of V_{DS} causes current to flow in the induced channel from Drain to Source. This is caused by the electrons present in the channel, and hence the current is only due to majority charge carriers in the channel.



Note that the application of V_{DS} will reverse bias the junction between Drain and Body, and the depletion region there will increase.

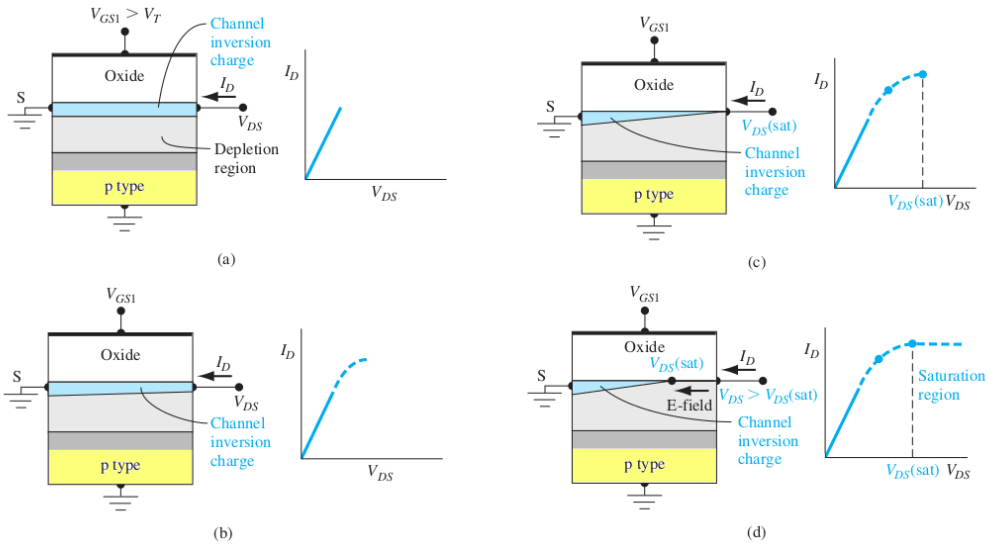
In linear region, the current has the given relationship with the applied voltages V_{DS} and V_{GS}.

$$I_D = \frac{\mu_n C_{ox} W}{L}\left((V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right)$$

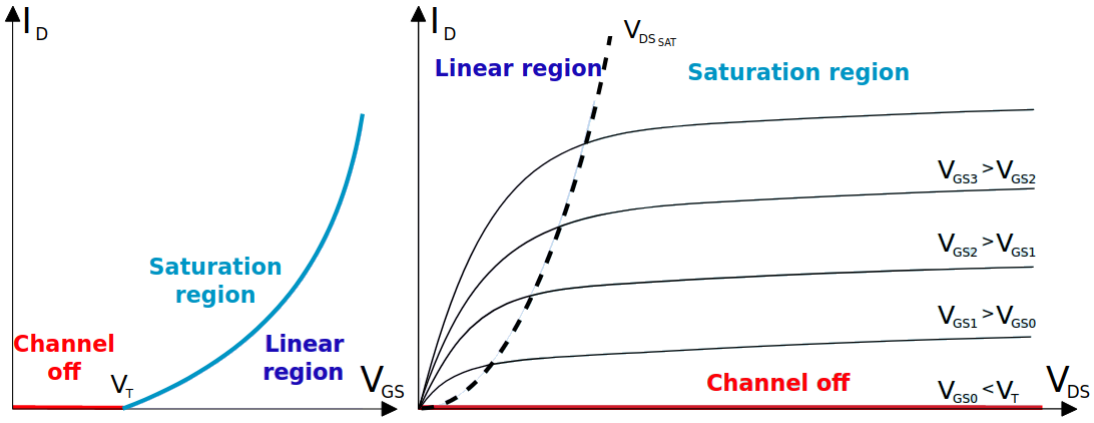
Here, μ_n is mobility of electrons, C_{ox} is the oxide capacitance, $\frac{W}{L}$ is the ratio of channel width to channel length and V_T is the threshold voltage.
This mode is used as 'on' state of a switch.

Saturation mode:
As V_{DS} is increased further, it reaches a point where any further increase will not affect the current. This is because higher values of V_{DS} will cause depletion region near the Drain to occupy the channel till it causes 'pinch-off' i.e a state when channel width can't decrease further and only very small region is available for current flow. This will ensure no further current can flow and hence the dependence of current I_D on V_{DS} is ideally lost. Quantitatively, this occurs when the value of V_{DS} goes beyond $V_{GS} - V_T$ (use the substitution $V_{DS_{sat}} = V_{GS} - V_T$ to find saturation current).
The relation between $I_{D_{sat}}$ and V_{GS} is hence given by,

$$I_{D_{sat}} = \frac{\mu_n C_{ox} W}{2L}(V_{GS} - V_T)^2$$



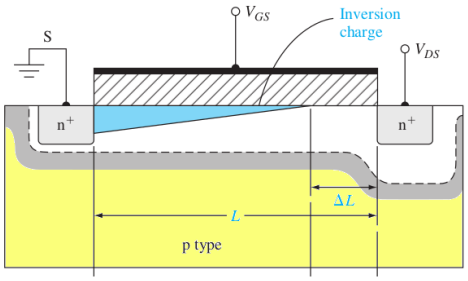
Transfer characteristics: I_D *vs* V_{GS}
Drain/Output characteristics: I_D *vs* V_{DS}



3.3.2 Channel length modulation

A non-ideality occurs when V_{DS} is increased beyond $V_{DS_{sat}}$ i.e in the saturation region, meaning there actually will be some dependency of V_{DS} on I_D.

As mentioned, high V_{DS} will cause depletion region near the Drain to occupy the channel till it causes 'pinch-off'. But as it is increased further, the depletion region around the Drain will expand enough to decrease the channel length.
This means, since channel length effectively decreases, the Drain current increases.



The modified expression for Drain current in saturation mode is given by,

$$I_D = \frac{\mu_n C_{ox} W}{2L}(V_{GS} - V_T)^2(1 + \lambda V_{DS})$$

Here, λ is called the "Channel length modulation" factor.

3.3.3 Other non-idealities

The most significant non-ideality is CLM. However, there are other non-idealities that may need to be considered in some cases, so knowing them is useful.

Sub-threshold conduction:

Ideally, it is assumed that when $V_{GS} < V_T$, no inversion has occurred and so $I_D = 0$ for any applied V_{DS}.
However, in reality the inversion phenomenon starts as soon as flat band voltage is achieved (as explained in MOS-Cap) i.e $V_{GS} > V_{FB}$. Hence, there will be some small current for Gate voltages $V_{FB} < V_{GS} < V_T$. This is called sub-threshold conduction.

In the sub-threshold mode, weak inversion would have occurred meaning channel is not fully formed but electrons are present. And electron concentration will be higher near Source than near Drain, so concentration gradient exists, which will cause diffusion current from Drain to Source even when V_{DS} is not applied.

Trapped charges:

Ideally, the oxide layer (insulator) has to be a perfect insulator but there are possibilities of trapped charges being present within the oxide layer due to manufacturing defects such as mobile ions or oxide charges or inter-facial trapped charges.
These trapped charges will affect the flat band voltage because they will also participate in compensating for the voltage applied to achieve flat band.

$$\implies V_{FB} = \phi_m - \phi_s - \gamma \frac{Q_{trap}}{C_{ox}}$$

Here, Q_{trap} sums up the total trapped charges present and γ is the relative average position of the charges inside the oxide layer compared to the oxide thickness.
If $\gamma \approx 1$, then charges are prominent near the oxide-semiconductor interface and if $\gamma \approx 0$, then charges are prominent near the oxide-metal interface.

Trapped charges also affects the threshold voltage V_T since they affect V_{FB}.

Dependence of V_T on depletion width:

Since the depletion width keeps changing with different applied voltages, the depletion charge is not a constant value. From the basic expression for threshold voltage V_T, it is obvious that it changes directly with Q_{dep} (which varies directly with x_d).
Hence, the threshold voltage itself is actually a function of applied voltages and won't be a constant value.

Body bias:

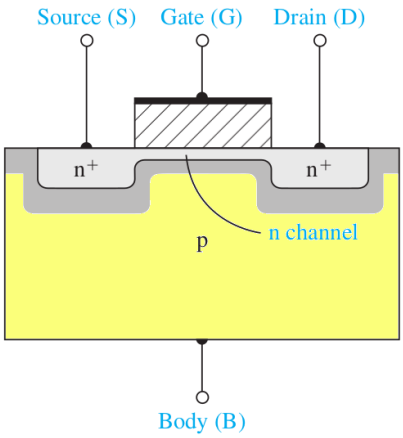
Earlier it was mentioned that the Body contact (substrate) is grounded. However, if some potential is provided to the Body, it will affect the threshold voltage.
This is because if $V_B \neq 0$, then the junction between Source and Body will be biased.

- $V_B < 0 \implies$ Reverse biased: Electrons migrate towards source, causing inversion to occur at lower built in potential $V_{bi} = 2\phi_f - |V_B|$, causing V_T to decrease.
- $V_B > 0 \implies$ Forward biased: Electrons migrate away from source, causing inversion to occur at higher built in potential $V_{bi} = 2\phi_f + |V_B|$, causing V_T to increase.

It can be noted that the working of a p-channel enhancement type MOSFET (pMOS) would have very similar working. The only difference is polarities of voltages required to bias a pMOS are reverse (and V_T will be negative as mentioned before), so all equations can be obtained simply by using V_{SG} and V_{SD} instead of V_{GS} and V_{DS} respectively.

3.3.4 Depletion type MOSFET

Unlike the enhancement type MOSFET studied earlier, in a depletion type MOSFET, the channel already exists.
The working of an n-channel depletion type MOSFET can be explained taking 3 different cases for the Gate voltage.

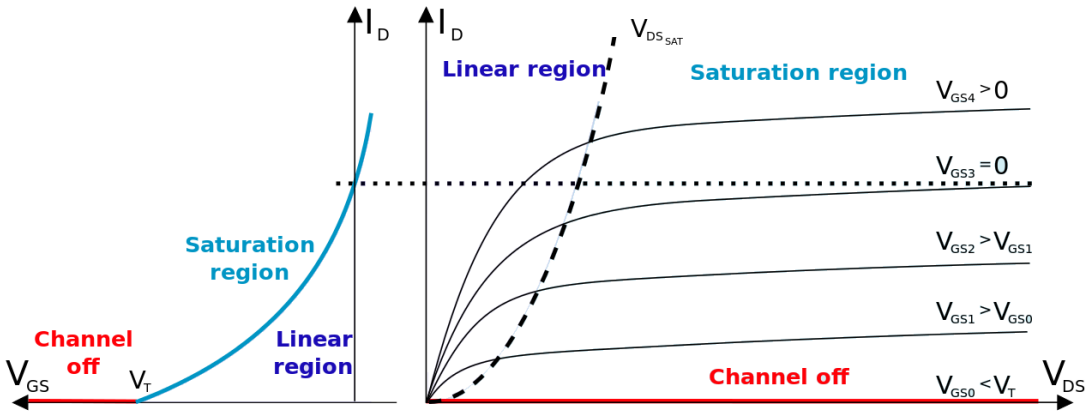


1. $V_{GS} = 0$:
   At zero Gate voltage, the applied Drain voltage V_{DS} causes current flow from Drain to Source due to presence of inbuilt channel.
   At lower values of V_{DS}, the relationship between I_D and V_{DS} is linear but at higher values it saturates for same reason as enhancement type MOSFET.

2. $V_{GS} < 0$:
   When negative Gate voltage is applied, the channel thickness reduces due to recombination in the channel. Eventually for more negative values, it ceases to exist and no current will flow.
   This is called "depletion mode".

3. $V_{GS} > 0$:
   When positive Gate voltage is applied, the number of free electrons in the channel increases. This causes more current to flow for the same V_{DS}. Hence, higher positive values of V_{GS} will increase current flow (till saturation is reached).
   This is called "enhancement mode".

The Transfer and Output characteristics depletion nMOS are shown.



The analogies between nMOS and pMOS hold for depletion type MOSFETs as well.