Basic Physics of Semiconductors

* Einstein Relation: Our study of drift and diffusion has introduced a factor for each:
\[ \mu_n \text{ (or } \mu_p \text{), and } D_n \text{ (or } D_p \text{), respectively. It can be proved that } \mu \text{ and } D \text{ are called the “Einstein Relation”:} \]
\[ \frac{D}{\mu} = \frac{kT}{q} \]
Note that: \( kT/q \approx 26 \text{ mV} \), at 300 K. Figure 1 summarizes the charge transport mechanisms.

\[ J_n = -q \mu_n n \nu n E \]
\[ J_p = -q \mu_p p \nu p E \]

Fig.1 Summary of drift and diffusion mechanisms

* PN Junction: We begin our study of semiconductor devices with the junction for three reasons. (1) The device finds application in many electronic systems, e.g., in adapters that charge the batteries of cell phones. (2) The \textit{pn} junction is among the simplest semiconductor devices, thus providing a good entry point into the study of the operation of such complex structures as transistors. (3) The \textit{pn} junction also serves as part of transistors.

In this section, we study the properties and I/V characteristics of \textit{pn} junctions. The following outline shows our thought process, indicating that our objective is to develop \textit{circuit} models that can be used in analysis and design. In a “\textit{pn} junction see Fig. 2, the \textit{p} and \textit{n} sides are called the “anode” and the ”cathode” respectively.
* PN Junction in *Equilibrium*

Let us first study the *pn* junction with no external connections, i.e., the terminals are open and no voltage is applied across the device. We say the junction is in “*equilibrium*”. While seemingly of no practical value, this condition provides insights that prove useful in understanding the operation under *no equilibrium* as well.

**Example 1:**

A *pn* junction employs the following doping levels: \( N_d = 10^{16} \text{ cm}^{-3} \) and \( N_A = 5 \times 10^{15} \text{ cm}^{-3} \).

Determine the hole and electron concentrations on the two sides.

**Solution:** We express the concentrations of holes and electrons on the *p* side respectively as:

\[
\begin{align*}
p_p & \approx N_d = 10^{16} \text{ cm}^{-3} \\
n_p & \approx \frac{n_i^2}{N_d} = \frac{(1.08 \times 10^{10} \text{ cm}^{-2})^2}{10^{16} \text{ cm}^{-3}} \\
 & \approx 1.1 \times 10^4 \text{ cm}^{-2}.
\end{align*}
\]

Similarly, the concentrations on the *n* side are given by:

\[
\begin{align*}
n_n & \approx N_d = 5 \times 10^{15} \text{ cm}^{-3} \\
p_n & \approx \frac{n_i^2}{N_d} = \frac{(1.08 \times 10^{10} \text{ cm}^{-2})^2}{5 \times 10^{15} \text{ cm}^{-3}} \\
 & \approx 2.4 \times 10^4 \text{ cm}^{-2}.
\end{align*}
\]

Note that the majority carrier concentration on each side is many orders of magnitude higher than the minority carrier concentration on either side.
* **Built-in Potential:** The existence of an electric field within the depletion region suggests that the junction may exhibit a “built-in potential”. Since the electric field \( E = -\frac{dV}{dx} \)

\[
q\mu_p E = qD_p \frac{d\rho}{dx}
\]

We have

\[
-\mu_p \frac{dV}{dx} = D_p \frac{d\rho}{dx}
\]

Dividing both sides by \( p \) and taking the integral, we obtain

\[
-\mu_p \int_{x_1}^{x_2} dV = D_p \int_{p_n}^{p_p} \frac{d\rho}{\rho}
\]

![Fig.3 Carrier profiles in a pn junction.](image)

Where \( p_n \) and \( p_p \) are the hole concentrations at \( x_1 \) and \( x_2 \) respectively.

\[
V(x_2) - V(x_1) = \frac{D_p}{\mu_p} \ln \frac{p_p}{p_n}
\]

The right side represents the voltage difference developed across the depletion region and will be denoted by \( V_0 \). Also, from Einstein’s relation, we can replace \( \mu_p \) with \( \frac{kT}{q} \):

\[
V_0 = \frac{kT}{q} \ln \frac{p_p}{p_n}
\]

**Exercise:**

Writing for electron drift and diffusion currents equation, and the integration, derive an equation for \( V_0 \) in term of \( n_n \) and \( n_p \).

Finally, for \( n_p \) and \( n_n \) yields

\[
V_0 = \frac{kT}{q} \ln \left( \frac{N_n N_p}{n_i^2} \right)
\]
Expressing the built-in potential in terms of junction parameters, this equation plays a central role in many semiconductor devices.

**Example 2:** A silicon \(pn\) junction employs \(N_a = 2 \times 10^{16} \text{ cm}^{-3}\) and \(N_d = 4 \times 10^{16} \text{ cm}^{-3}\). Determine the built-in potential at room temperature (\(T=300 \text{ K}\)).

**Solution:**

Since silicon intrinsic concentration \(n_i\), thus,

\[
V_o \approx \frac{(26 \text{ mV}) \ln \left( \frac{(2 \times 10^{16}) \times (4 \times 10^{16})}{(1.68 \times 10^{10})^2} \right)}{\left(1.68 \times 10^{10}\right)^2} \approx 768 \text{ mV}
\]

**Example 3:** Reveals that \(V_o\) is a weak function of the doping levels. How much does \(V_o\) change if \(N_a\) or \(N_d\) is increased by one order of magnitude?

\[
\frac{V_o}{V_T} = \frac{kT}{q} = 26 \text{ mV}, \quad \frac{V_o}{V_T} = V_T \ln(10) \frac{n_i^2}{N_a N_d} - V_T \ln(N_a N_d)
\]

Let

\[
V_o = V_T \ln 10
\]

An interesting question may arise at this point. The junction carries no net current (because its terminals remain open), but it sustains a voltage. How is that possible? We observe that the built-in potential is developed to oppose the flow of diffusion currents (and is, in fact, sometimes called the “potential barrier.”).

* **PN Junction Under Reverse Bias:** For the case of reverse bias, we note that in Fig. 4, as \(V_R\) increases, more positive charge appears on the n side and more negative charge on the p side.
Thus, the device operates as a capacitor [Fig. 5(a)]. In essence, we can view the conductive n and p sections as the two plates of the capacitor. We also assume the charge in the depletion region equivalently resides on each plate.

Since any two parallel plates can form a capacitor, for increasing values of $V_R$, revealing that the capacitance of the structure decreases as the two plates move away from each other. The junction therefore displays a voltage-dependent capacitance.

It can be proved that the capacitance of the junction per unit area is equal to:

$$C_J = \frac{C_{jo}}{\sqrt{1 - \frac{V_R}{V_o}}}$$

$c_{jo}$ denotes the capacitance corresponding to zero bias ($V_R = 0$) and $V_o$ is the built-in potential.


\[ C_{j_0} = \frac{\varepsilon \varepsilon_0 N_a N_d}{2 N_a + N_d V_0} \]

Where \( \varepsilon_0 \) represents the dielectric constant of silicon and is equal to

The dielectric constant of materials is usually written in the form \( \varepsilon_0 \varepsilon_r \) where \( \varepsilon_r \) is the “relative”
dielectric constant and a dimensionless factor (e.g., 11.7), and \( \varepsilon_0 \) the dielectric constant of
vacuum (8.85 \( 10^{-14} \) F/cm). Fig. 6 shows \( C_j \) decreases as \( V_R \) increases.

Fig.6 Junction capacitance under reverse bias

**Example 4**: A pn junction is doped with \( N_a = 2 \times 10^{16} \text{cm}^{-2} \) and \( N_d = 9 \times 10^{15} \text{cm}^{-2} \).
Determine the capacitance of the device with (a) \( V_R = 0 \) and \( V_R = 1 \text{ V} \).

**Solution**: We first obtain the built-in potential:

\[
V_0 = V_T \ln \left( \frac{N_a N_d}{n_i^2} \right)
\]

Thus, for \( V_R = 0 \) and \( q = 1.6 \times 10^{-29} \text{ C} \), we have

\[
C_{j_0} = \frac{\varepsilon \varepsilon_0 N_a N_d}{2 N_a + N_d V_0} = 2.65 \times 10^{-9} \frac{\text{F}}{\text{cm}^2}.
\]

In microelectronics, we deal with very small devices and may rewrite this result as

\[
C_{j_0} = 0.265 \frac{F}{\mu \text{m}^2}.
\]

where, 1fF (femtofarad) = \( 10^{-15} \) F, for \( V_R = 1 \text{ V} \)

\[
C_j = \frac{C_{j_0}}{\sqrt{1 - \frac{V_R}{V_0}}}
\]
\[ f_{\text{res}} = \frac{f_F}{\mu m^2} \]

**Example 5**: A cellphone incorporates a 2-GHz oscillator whose frequency is defined by the resonance frequency of an LC tank (Fig. 7). If the tank capacitance is realized as the \( pn \) junction of Example 4, calculate the change in the oscillation frequency while the reverse voltage goes from 0 to 2 V. Assume the circuit operates at 2 GHz at a reverse voltage of 0 V, and the junction area is 2000 \( \mu m^2 \).

Fig. 7 Variable capacitor used to tune an oscillator.

Recall from basic circuit theory that the tank “resonates” if the impedances of the inductor and the capacitor are equal and opposite: \( jL\omega_{\text{res}} = (jC\omega_{\text{res}})^2 \). Thus, the resonance frequency \( (f_{\text{res}}) \) is equal to

\[ f_{\text{res}} = \frac{1}{2\pi\sqrt{LC}} \]

At \( V_R = 0 \), \( C_j = 0.265 \frac{f_F}{\mu m^2} \), yielding a total device capacitance of

\[ C_{j,\text{tot}} = \left(0.265 \frac{f_F}{\mu m^2}\right) \times (2000 \mu m^2) = 530 fF \]

Setting \( f_{\text{res}} \) to 2 GHz, we obtain

\[ L = 11.9 nH \]

If \( V_R \) goes to 2V

\[ C_{j,\text{tot}}(V_R = 2V) = \frac{C_j}{\sqrt{1 - \frac{2}{0.73}}} \times 2000 \mu m^2 = 274 fF \]

Using this value along with \( L = 11.9 nH \), we have

\[ f_{\text{res}}(V_R = 2V) = 2.79 \text{ GHz} \]
An oscillator whose frequency can be varied by an external voltage \( V \text{ in this case} \) is called a “voltage-controlled oscillator” and used extensively in cell phones, microprocessors, personal computers, etc.

**Exercise:** Some wireless systems operate at 5.2 GHz. Repeat the above example for this frequency, assuming the junction area is still 2000 \( \mu \text{m}^2 \) but the inductor value is scaled to reach 5.2 GHz.

**In summary:**
A reverse-biased \( pn \) junction carries a *negligible current* but exhibits a *voltage dependent* capacitance. Note that we have tacitly developed a circuit model for the device under this condition: a simple capacitance \( C_j \).

Another interesting application of reverse-biased diodes is in digital cameras. If light of sufficient energy is applied to a \( pn \) junction, electrons are dislodged from their covalent bonds and hence electron-hole pairs are created. With a reverse bias, the electrons are attracted to the positive battery terminal and the holes to the negative battery terminal. As a result, current flows through the diode that is proportional to the light intensity. We say the \( pn \) junction operates as a *photodiode."

* I/V Characteristics:
Let us summarize our thoughts thus far. In forward bias, the external voltage opposes the built-in potential, raising the diffusion currents substantially. In reverse bias, on the other hand, the applied voltage enhances the field, prohibiting current flow. We hereafter write the junction equation as:

\[
I_D = I_S \left( \frac{\exp{\frac{V_D}{V_T}} - 1}{1} \right) \quad (1)
\]

Where \( I_D \) and \( V_D \) denote the diode current and voltage, respectively. As expected \( V_D = 0 \) yields \( I_D = 0 \) (Why is this expected?) As \( V_D \) becomes positive and exceeds several \( V_T \) the exponential
term grows rapidly and 

\[ I_D \approx I_S \exp \left( \frac{V_D}{V_T} \right) \]. We hereafter assume 

\[ \exp \left( \frac{V_D}{V_T} \right) \gg 1 \]

in the forward bias region.

It can be proved that Eq.1 also holds in reverse bias, i.e., for negative \( V_D \). If \( V_D < 0 \) and \( |V_D| \) reaches several \( V_T \), then

\[ \exp \left( \frac{V_D}{V_T} \right) \ll 1 \]

and

\[ I_D \approx -I_S \]

\[ I_S = A q \alpha \left( \frac{D_n}{N_a L_n} + \frac{D_p}{N_d L_p} \right) \] \text{.........(2)}

In equation (2), \( A \) is the cross section area of the device, and \( L_n \) and \( L_p \) are electron and hole “diffusion lengths,” respectively. Diffusion lengths are typically in the range of tens of micrometers. Note that the first and second terms in the parentheses correspond to the flow of electrons and holes, respectively.

**Example 6:**

Determine \( I_S \) for the junction of Example 2 at \( T = 300 \) K if \( L_n = 20 \) \( \mu \)m, and

\[ D_n = 3 \times 10^{-5} \text{ cm}^2/\text{s}, \quad n_i = 1.0 \times 10^{10} \text{ electron/cm}^3, \quad D_p = 3 \times 10^{-5} \text{ cm}^2/\text{s}, \quad \text{and} \]

\[ D_p = 12 \text{ cm}^2/\text{s}, \quad \text{we have} \]

\[ I_S = 1.77 \times 10^{-17} \text{ A}. \]

**Exercise:** What junction area is necessary to raise \( I_S \) to \( 10^{-15} \) A.

Figure 8 plots the overall I/V characteristic of the junction, revealing why \( I_S \) is called the “reverse saturation current.” Example 6 indicates that \( I_S \) is typically very small. We therefore
view the current under reverse bias as “leakage.” Note that $I_S$ and hence the junction current is proportional to the device cross section area [see equation (2)]. For example, two identical devices placed in parallel (Fig. 9) behave as a single junction with twice the $I_S$.

**Example 7:**
Each junction in Fig. 9 employs the doping levels described in Example 2. Determine the forward bias current of the composite device for $V_D = 300$ mV and 800 mV at $T = 300$ K.

**Solution:**
From example 6, $I_S = 1.77 \times 10^{-27} \text{A}$, for each junction. Thus, the total current is equal to

Similarly, for $V_D = 800$ mV:

$I_{D,\text{for}}(V_D = 300 \text{ mV}) = 0.2 \mu\text{A}$. 

Fig. 8 I-V characteristic of a $pn$ junction

Fig. 9 Equivalence of parallel devices to a larger device
Exercise: How many of these diodes must be placed in parallel to obtain a current of 100 \(\mu A\) with a voltage of 750 mV.

Example 8: A diode operates in the forward bias region with a typical current level [i.e., \(I_D = I_S \exp\left(\frac{V_D}{V_T}\right)\)]. Suppose we wish to increase the current by a factor of 10. How much change in \(V_D\) is required?

Let us first express the diode voltage as a function of its current:

\[
V_D = \frac{V_T \ln I_D}{I_S}
\]

We define \(i_1 = 10 I_D\) and seek the corresponding voltage, \(V_{D1}\):

\[
V_{D1} = \frac{V_T \ln 10 I_D}{I_S} = V_T \ln I_D + V_T \ln 10
\]

Thus, the diode voltage must rise by \(V_T \ln 10 \approx 60 \text{ mV}\) (at \(T=300\text{K}\)) to accommodate a tenfold increase in the current. We say the device exhibits a 60-mV/decade characteristic, meaning \(V_D\) changes by 60 mV for a decade (tenfold) change in \(I_D\). More generally, an \(n\)-fold change in \(I_D\) translates to a change of \(V_T \ln n\) in \(V_D\).

Exercise: By what factor does the current change if the voltages changes by 120 mV?

Example 9: The cross section area of a diode operating in the forward bias region is increased by a factor of 10. (a) Determine the change in \(I_D\) if \(V_D\) is maintained constant. (b) Determine the change in \(V_D\) if \(I_D\) is maintained constant. Assume

\[
I_{D1} = 10 I_D \exp\left(\frac{V_D}{V_T}\right)
\]

\[
= 10 I_D
\]

(b) From the above example,

\[
\frac{V_{D1} - V_T \ln I_D}{10 I_S} = V_T \ln 10
\]
Thus, a tenfold increase in the device area lowers the voltage by 60 mV if $I_D$ remains constant.

**Exercise:** A diode in forward bias with $I_D \approx I_S \exp \left( \frac{V_D}{V_T} \right)$ undergoes two simultaneous changes: the current is raised by a factor of $m$ and the area is increased by a factor of $n$. Determine the change in the device voltage.

*Reverse Breakdown:* Recall from Fig. 8 that the $pn$ junction carries only a small, relatively constant current in reverse bias. However, as the reverse voltage across the device increases, eventually “breakdown” occurs and a sudden, enormous current is observed. Figure 10 plots the device I/V characteristic, displaying this effect.

The breakdown resulting from a high voltage (and hence a high electric field) can occur in any material. A common example is lightning, in which case the electric field in the air reaches such a high level as to ionize the oxygen molecules, thus lowering the resistance of the air and creating a tremendous current.

The breakdown phenomenon in $pn$ junctions occurs by one of two possible mechanisms:

"Zener effect" and "avalanche effect."

1. **Zener Breakdown**

The depletion region in a $pn$ junction contains atoms that have lost an electron or a hole and, therefore, provide no loosely-connected carriers (no free carriers). However, a high electric field in this region may impart enough energy to the remaining covalent electrons to tear them from their bonds [Fig. 11(a)]. Once freed, the electrons are accelerated by the field and swept to the $n$ side of the junction. This effect occurs at field strength of about $10^8 \frac{V}{cm}$.
In order to create such high fields with reasonable voltages, a narrow depletion region is required, which from Eq. of \( \frac{C_{fe}}{C_{je}} \) translates to high doping levels on both sides of the junction (why?). Called the “Zener effect,” this type of breakdown appears for reverse bias voltages on the order of 3-8 V.

![Fig.11](attachment:image.png) (a) Release of electrons due to high electric field, (b) avalanche effect

### 2. Avalanche Breakdown

Junctions with moderate or low doping levels \( \left( < 10^{15} \text{ cm}^{-2} \right) \) generally exhibit no Zener breakdown. But, as the reverse bias voltage across such devices increases, an avalanche effect takes place. Even though the leakage current is very small, each carrier entering the depletion region experiences a very high electric field and hence a large acceleration, thus gaining enough energy to break the electrons from their covalent bonds. Called “impact ionization,” this phenomenon can lead to avalanche: each electron freed by the impact may itself speed up so much in the field as to collide with another atom with sufficient energy, thereby freeing one more covalent-bond electron. Now, these two electrons may again acquire energy and cause more ionizing collisions, rapidly raising the number of free carriers.

An interesting contrast between Zener and avalanche phenomena is that they display opposite temperature coefficients (TCs): \( V_{BD} \) has a negative TC for Zener effect and positive TC for avalanche effect. The two TCs cancel each other for \( V_{BD} = 3.5 \text{ V} \). For this reason, Zener diodes with 3.5-V rating find application in some voltage regulators.
The Zener and avalanche breakdown effects do not damage the diodes if the resulting current remains below a certain limit given by the doping levels and the geometry of the junction. Both the breakdown voltage and the maximum allowable reverse current are specified by diode manufacturers.

**Summary**

- Silicon contains four atoms in its last orbital. It also contains a small number of free electrons at room temperature.
- When an electron is freed from a covalent bond, a “hole” is left behind.
- The bandgap energy is the minimum energy required to dislodge an electron from its covalent bond.
- To increase the number of free carriers, semiconductors are “doped” with certain impurities. For example, addition of phosphorous to silicon increases the number of free electrons because phosphorous contains five electrons in its last orbital.
- For doped or undoped semiconductors $p^n = n^p$. For example, in an $n$-type material, $n \approx N_d$ and hence $p \approx \frac{n^p}{N_d}$.
- Charge carriers move in semiconductors via two mechanisms: drift and diffusion.
- The drift current density is proportional to the electric field and the mobility of the carriers and is given by $I_{\text{drift}} = q(\mu_n n + \mu_p p)E$.
- The diffusion current density is proportional to the gradient of the carrier concentration and given by $I_{\text{diff}} = q\left(D_n \frac{dn}{dx} + D_p \frac{dp}{dx}\right)$.
- A $pn$ junction is a piece of semiconductor that receives $n$-type doping in one section and $p$-type doping in an adjacent section.
- The $pn$ junction can be considered in three modes: equilibrium, reverse bias, and forward bias.
- Upon formation of the $pn$ junction, sharp gradients of carrier densities across the junction result in a high current of electrons and holes. As the carriers cross, they leave ionized atoms behind, and a “depletion region” is formed. The electric field created in the depletion region eventually stops the current flow. This condition is called equilibrium.
• The electric field in the depletion results in a built-in potential across the region equal
to

\[ V_0 = \frac{kT}{q} \ln \left( \frac{N_a N_d}{n_i^2} \right) \]

, typically in the range of 700 to 800 mV.

• Under reverse bias, the junction carries negligible current and operates as a capacitor. The capacitance itself is a function of the voltage applied across the device.

• Under forward bias, the junction carries a current that is an exponential function of the applied voltage:

\[ I_S \exp \left( \frac{V_F}{V_T} - 1 \right) \]

Since the exponential model often makes the analysis of circuits difficult, a constant voltage model may be used in some cases to estimate the circuit’s response with less mathematical labor.

• Under a high reverse bias voltage, pn junctions break down, conducting a very high current. Depending on the structure and doping levels of the device, “Zener” or “avalanche” breakdown may occur.
**Physics of Bipolar Junction Transistors (BJT)**

The bipolar transistor was invented in 1945 by Shockley, subsequently replacing vacuum tubes in electronic systems and paving the way for integrated circuits.

**Structure of Bipolar Transistor**

The bipolar transistor consists of three doped regions forming a sandwich. Fig. 1 (a) is an example comprising of a $P$ layer sandwiched between two $n$ regions and called an “$npi$” transistor. The three terminals are called the “base,” the “emitter,” and the “collector”. The emitter “emits” charge carriers and the collector “collects” them while the base controls the number of carriers that make this journey. The circuit symbol for the “$npi$” transistor is shown in Fig. 1 (b). We denote the terminal voltages by $V_E, V_B$, and $V_C$, and the voltage differences by $V_{BE}$.

The transistor is labeled $Q_1$ here, the device contains two $pn$ junction diodes.

![Fig. 1 (a) Structure and (b) circuit symbol of bipolar transistor](image)

For example, if the base is more positive than the emitter, $V_{BE} > 0$, then this junction is forward-biased. $E$ and $C$ (the emitter and the collector), in reality, cannot be interchanged.
because the dimensions and doping levels of these two regions are quite different. Also a proper operation requires a thin base region, e.g., about 100 Å in modern integrated bipolar transistors.

**Operation of Bipolar Transistor in Active Mode**

To act as a voltage-controlled current source, more specifically, we intend to show that (1) the current flow from the emitter to the collector can be viewed as a current source tied between these two terminals, and (2) this current is controlled by the voltage difference between the base and the emitter, $V_{BE}$. We begin our study with the assumption that the base-emitter junction is forward-biased ($V_{BE} > 0$) and the base-collector junction is reverse-biased ($V_{BC} < 0$). Under these conditions, we say the device is biased in the “forward active region” or simply in the “active mode.”

For example: with the emitter connected to ground, the base voltage is set to about 0.8 V and the collector voltage to a higher value, e.g., 1 V [Fig.2 (a)]. The base-collector junction therefore experiences a reverse bias of 0.2 V.

![Fig.2 (a) Bipolar transistor with base and collector bias voltages, (b) Simplistic view of bipolar transistor.](image)

Examine the flow of charge inside the device, bearing in mind that the base region is very thin. Since the EB (base-emitter) junction is forward-biased, electrons flow from the emitter to the base and holes from the base to the emitter. For proper transistor operation, emitter current component must be much greater than base, requiring that the emitter doping level be much greater than that of the base. Thus, we denote the emitter region with $n^+$, where the superscript emphasizes the high doping level. Fig.3 (a) summarizes our observations thus far, indicating that
the emitter injects a large number of electrons into the base while receiving small number of holes from it.

What happens to electrons as they enter the base? Since the base region is thin, most of the electrons reach the edge of the collector-base depletion region, beginning to experience the built-in electric field. Consequently, the electrons are swept into the collector region and absorbed by the positive battery terminal. Fig.3 (b) shows that the reverse-biased collector base junction carries a current because minority carriers are “injected” into its depletion region.

![Diagrams showing flow of electrons and holes through base-emitter junctions.](image)

Let us summarize our thoughts. In the active mode, an $n\!p\!n$ bipolar transistor carries a large number of electrons from the emitter, through the base, to the collector while drawing a small current of holes through the base terminal. We must now answer several questions. First, how do electrons travel through the base: by drift or diffusion? Second, how does the resulting current depend on the terminal voltages? Third, how large is the base current?

Operating as a moderate conductor, the base region sustains but a small electric field, i.e., it allows most of the field to drop across the base-emitter depletion layer. Thus, as explained for $p\!n$ junctions in the previous lecture, the drift current in the base is negligible, leaving diffusion as
the principal mechanism for the flow of electrons injected by the emitter. In fact, two observations directly lead to the necessity of diffusion: that the density of electrons at the emitter base junction is very high.

**Summary: (Physics of Bipolar Junction Transistors (BJT))**

- Bipolar transistors are electronic devices that can operate as voltage-dependent current sources.
- The bipolar transistor consists of two $pn$ junctions and three terminals: base, emitter, and collector. The carriers flow from the emitter to the collector and are controlled by the base.
- For proper operation, the base-emitter junction is forward-biased and the base-collector junction reverse-biased (forward active region). Carriers injected by the emitter into the base approach the edge of collector depletion region and are swept away by the high electric field.
- The large-signal model of the bipolar transistor consists of an exponential voltage-dependent current source tied between the collector and emitter, and a diode (accounting for the base current) tied between the base and emitter.
- The small-signal model of bipolar transistors consists of a linear voltage-dependent current source, a resistance tied between the base and emitter, and an output resistance.
- If the base-collector junction is forward-biased, the bipolar transistor enters saturation and its performance degrades.
- The small-signal models of $nnp$ and $pnp$ transistors are identical.
**Physics of MOS Transistors**

Today’s field of microelectronics is dominated by a type of device called the metal-oxide-semiconductor field-effect transistor (MOSFET). Conceived in the 1930s but first realized in the 1960s, MOSFETs (also called MOS devices) offer unique properties that have led to the revolution of the semiconductor industry. This revolution has culminated in microprocessors having 100 million transistors, memory chips containing billions of transistors, and sophisticated communication circuits providing tremendous signal processing capability.

1. **Structure of MOSFET**

In order to arrive at the structure of the MOSFET, we begin with a simple geometry consisting of a conductive (e.g., metal) plate, an insulator (“dielectric”), and a doped piece of silicon. Illustrated in Fig.1 (a), such a structure operates as a capacitor because the $p$-type silicon is somewhat conductive, “mirroring” any charge deposited on the top plate.

If a potential difference is applied: As positive charge is placed on the top plate as shown in Fig.1 (b), it attracts negative charge, e.g., electrons, from the piece of silicon. (Even though doped with acceptors, the $p$-type silicon does contain a small number of electrons.) We therefore observe that a “channel” of *free* electrons may be created at the interface between the insulator and the piece of silicon, potentially serving as a good conductive path if the electron density is sufficiently high. The key point here is that the density of electrons in the channel varies with $V_1$, as evident from $Q = CV$, where $C$ denotes the capacitance between the two plates.

---

![Diagram](image.png)

Fig 1 (a) Hypothetical semiconductor device, (b) operation as a capacitor, (c) current flow as a result of potential difference
The dependence of the electron density upon $V_x$ leads to an interesting property: if, as depicted in Fig.1 (c), we allow a current to flow from left to right through the silicon material, $V_x$ can control the current by adjusting the resistivity of the channel. (Note that the current prefers to take the path of least resistance, thus flowing primarily through the channel rather than through the entire body of silicon.) This will serve our objective of building a voltage-controlled current source.

Equation suggests that, to achieve a strong control of $V_x$ by $V$, the value of $C$ must be maximized, for example, by reducing the thickness of the dielectric layer separating the two plates. The capacitance between two plates is given by $\frac{\varepsilon A}{t}$, where $\varepsilon$ is the “dielectric constant” (also called the “permittivity”), $A$ is the area of each plate, and $t$ is the dielectric thickness.

The “gate” (G), the top conductive plate resides on a thin dielectric (insulator) layer, which itself is deposited on the underlying $p$-type silicon “substrate.” To allow current flow through the silicon material, two contacts are attached to the substrate through two heavily-doped $n$-type regions because direct connection of metal to the substrate would not produce a good “ohmic” contact. These two terminals are called “source” (S) and “drain” (D) to indicate that the former can provide charge carriers and the latter can absorb them. This transistor operates with electrons rather than holes and is therefore called an $n$-type MOS (NMOS) device. We draw the device as shown in Fig.2 (b) for simplicity. Fig.2 (c) depicts the circuit symbol for an NMOS transistor, wherein the arrow signifies the source terminal.
The gate plate must serve as a good conductor and was in fact realized by metal (aluminum) in the early generations of MOS technology. However, it was discovered that non-crystalline silicon ("polysilicon" or simply "poly") with heavy doping (for low resistivity) exhibits better fabrication and physical properties. Thus, today’s MOSFETs employ polysilicon gates. The dielectric layer sandwiched between the gate and the substrate plays a critical role in the performance of transistors and is created by growing silicon dioxide (or simply “oxide”) on top of the silicon area.

The $n^+$ regions are sometimes called source/drain “diffusion.” We should also remark that these regions in fact form diodes with the $p$-type substrate (Fig.3). As explained later, proper operation of the transistor requires that these junctions remain reverse-biased. Thus, only the
depletion region capacitance associated with the two diodes must be taken into account. Fig.3 shows some of the device dimensions in today’s state-of-the-art MOS technologies. The oxide thickness is denoted by $\varepsilon_{ox}$.

Our study of the simple structures shown in Figs.1 and 2 suggests that the MOSFET may conduct current between the source and drain if a channel of electrons is created by making the gate voltage sufficiently positive. Moreover, we expect that the magnitude of the current can be controlled by the gate voltage.

Consider the arrangement shown in Fig.5 (a), where the source and drain are grounded and the gate voltage is varied. This circuit does not appear particularly useful but it gives us a great deal of insight. Recall from Fig.1 (b) that, as $V_G$ rises, the positive charge on the gate must be mirrored by negative charge in the substrate. While electrons are attracted to the interface, in reality, another phenomenon precedes the formation of the channel. As $V_G$ increases from zero, the positive charge on the gate repels the holes in the substrate, thereby exposing negative ions and creating a depletion region Fig.5 (b). [Note that: this depletion region contains only one immobile charge polarity, whereas the depletion region in a pn junction consists of two areas of negative and positive ions on the two sides of the junction]. Note that the device still acts as a capacitor—positive charge on the gate is mirrored by negative charge in the substrate—but no channel of mobile charge is created yet. Thus, no current can flow from the source to the drain. We say the MOSFET is off.

Can the source-substrate and drain-substrate junctions carry current in this mode? To avoid this effect, the substrate itself is also tied to zero, ensuring that these diodes are not forward biased. For simplicity, we do not show this connection in the diagrams.

What happens as $V_G$ increases? To mirror the charge on the gate, more negative ions are exposed and the depletion region under the oxide becomes deeper. Does this mean the transistor never turns on?! Fortunately, if $V_G$ becomes sufficiently positive, free electrons are attracted to the oxide-silicon interface, forming a conductive channel Fig.5 (c). We say the MOSFET is on. The gate potential at which the channel begins to appear is called the “threshold voltage,” $V_{TH}$ and falls in the range of 300 mV to 500 mV. Note that the electrons are readily provided by the $n^+$ source and drain regions, and need not be supplied by the substrate.
It is interesting to recognize that the gate terminal of the MOSFET draws no (low-frequency) current. Resting on top of the oxide, the gate remains insulated from other terminals and simply operates as a plate of a capacitor.

**MOSFET as a Variable Resistor** The conductive channel between S and D can be viewed as a resistor. Furthermore, since the density of electrons in the channel must increase as $V_G$ becomes more positive (why?), the value of this resistor changes with the gate voltage. Conceptually illustrated in Fig.6, such a voltage-dependent resistor proves extremely useful in analog and digital circuits.

**Example 1**

In the vicinity of a wireless base station, the signal received by a cell-phone may become very strong, possibly “saturating” the circuits and prohibiting proper operation. Devise a variable-gain circuit that lowers the signal level as the cell-phone approaches the base station.
Solution

A MOSFET can form a voltage-controlled attenuator along with a resistor as shown in Fig. 7.

Fig. 7 Use of MOSFET to adjust signal levels.

Since,

\[
\frac{v_{\text{out}}}{v_{\text{in}}} = \frac{R_1}{R_M + R_2}
\]

the output signal becomes smaller as \( V_{\text{cont}} \) falls because the density of electrons in the channel decreases and \( R_M \) rises. MOSFETs are commonly utilized as voltage-dependent resistors in “variable-gain amplifiers.”

Exercise

What happens to \( R_M \) if the channel length is doubled?

In the arrangement of Fig. 5 (c), no current flows between S and D because the two terminals are at the same potential. We now raise the drain voltage as shown in Fig. 8 (a) and examine the drain current (= source current). If \( V_G < V_{TH} \), no channel exists, the device is off, and \( I_D = 0 \) regardless of the value of \( V_D \). On the other hand, if \( V_G > V_{TH} \), then \( I_D > 0 \) [Fig. 8 (b)].

In fact, the source-drain path may act as a simple resistor, yielding the \( I_D = V_D \) characteristic shown in Fig. 8 (c). The slope of the characteristic is equal to \( \frac{I}{R_{\text{on}}} \), where \( R_{\text{on}} \) denotes the: “on-resistance” of the transistor.

Note: The term “on-resistance” always refers to that between the source and drain as no resistance exists between the gate and other terminals.
Our brief treatment of the MOS I/V characteristics thus far points to two different views of the operation: in Fig.8 (b), $V_G$ is varied while $V_D$ remains constant whereas in Fig.8 (c), $V_D$ is varied while $V_G$ remains constant. Each view provides valuable insight into the operation of the transistor.

How does the characteristic of Fig.8 (b) change if $V_G$ increases? The higher density of electrons in the channel lowers the on-resistance, yielding a greater slope. Depicted in Fig.8 (d), the resulting characteristics strengthen the notion of voltage-dependent resistance.

Knowing that charge flow in semiconductors occurs by diffusion or drift. How about the transport mechanism in a MOSFET? Since the voltage source tied to the drain creates an electric field along the channel, the current results from the drift of charge.

The $I_D - V_G$ and $I_D - V_D$ characteristics shown in Figs.8 (b) and (c), respectively, play a central role in our understanding of MOS devices. The following example reinforces the concepts studied thus far.
Example 2

Sketch the $I_D - V_G$ and $I_D - V_D$ characteristics for (a) different channel lengths, and (b) different oxide thicknesses.

Solution

As the channel length increases, so does the on-resistance (note that the resistance of a conductor is proportional to the length.) Thus, for $V_G > V_{TH}$, the drain current begins with lesser values as the channel length increases Fig.9 (a). Similarly, $I_D$ exhibits a smaller slope as a function of $V_D$ Fig.9 (b). It is therefore desirable to minimize the channel length so as to achieve large drain currents—an important trend in the MOS technology development.

How does the oxide thickness $t_{ox}$ affect the I-V characteristics? As $t_{ox}$ increases, the capacitance between the gate and the silicon substrate decreases. Thus, from $Q = CV$ we note that a given voltage results in less charge on the gate and hence a lower electron density in the channel. Consequently, the device suffers from a higher on-resistance, producing less drain current for a given gate voltage Fig.9 (c) or drain voltage Fig.9 (d). For this reason, the semiconductor industry has continued to reduce the gate oxide thickness.

Fig.9 (a) $I_D - V_G$ characteristics for different channel lengths, (b) $I_D - V_D$ characteristics for different channel lengths, (c) $I_D - V_G$ characteristics for different oxide thicknesses, (d) $I_D - V_D$ characteristics for different oxide thicknesses.
Exercise

The current conduction in the channel is in the form of drift. If the mobility falls at high temperatures, what can we say about the on-resistance as the temperature goes up?

While both the length and the oxide thickness affect the performance of MOSFETs, only the former is under the circuit designer’s control, i.e., it can be specified in the “layout” of the transistor. The latter, on the other hand, is defined during fabrication and remains constant for all transistors in a given generation of the technology.

Another MOS parameter controlled by circuit designers is the width of the transistor, the dimension perpendicular to the length [Fig.10 (a)]. We therefore observe that “lateral” dimensions such as $L$ and $W$ can be chosen by circuit designers whereas “vertical” dimensions such as $t_{ox}$ cannot.

How does the gate width impact the I-V characteristics? As $W$ increases, so does the width of the channel, thus lowering the resistance between the source and the drain and yielding the trends depicted in Fig.10 (b). From another perspective, a wider device can be viewed as two narrower transistors in parallel, producing a high drain current Fig.10 (c). We may then surmise that $W$ must be maximized, but we must also note that the total gate capacitance increases with $W$, possibly limiting the speed of the circuit. Thus, the width of each device in the circuit must be chosen carefully.

Note that the resistance of a conductor is inversely proportional to the cross section area, which itself is equal to the product of the width and thickness of the conductor.
Fig. 10 (a) Dimensions of a MOSFET ($W$ and $L$ are under circuit designer’s control.), (b) $I_D$ characteristics for different values of $W$ (c) equivalence to devices in parallel.

**Channel Pinch-Off** Our qualitative study of the MOSFET thus far implies that the device acts as a voltage-dependent resistor if the *gate voltage* exceeds $V_{TH}$. In reality, however, the transistor operates as a *current source* if the *drain voltage* is sufficiently positive. To understand this effect, we make two observations: (1) to form a channel, the potential difference between the gate and the oxide-silicon interface must exceed $V_{TH}$; (2) if the drain voltage remains higher than the source voltage, then the voltage at each point along the channel with respect to ground increases as we go from the source towards the drain. Illustrated in Fig. 11 (a), this effect arises from the gradual voltage drop along the channel resistance. Since the gate voltage is constant (because the gate is conductive but carries no current in any direction), and since the potential at the oxide-silicon interface rises from the source to the drain, the potential difference *between* the gate and the oxide-silicon interface *decreases* along the x-axis Fig. 11(b). The density of electrons in the channel follows the same trend, falling to a minimum at $x = L$.

![Diagram](image)

**Fig. 11** (a) Channel potential variation, (b) gate-substrate voltage difference along the channel
From these observations, we conclude that, if the drain voltage is high enough to produce $V_G - V_D \leq V_{TH}$, then the channel ceases to exist or (finished) near the drain. We say the gate-substrate potential difference is not sufficient at $x = L$ to attract electrons and the channel is “pinched off”.

Does this mean the transistor cannot conduct current? No, the device still conducts: once the electrons reach the end of the channel, they experience the high electric field in the depletion region surrounding the drain junction and are rapidly swept to the drain terminal, and the MOSFET acts as a constant current source—similar to a bipolar transistor in the forward active region. Note that the source-substrate and drain-substrate junctions carry no current.

**PMOS Transistor**

In MOSFETs a p-type counterpart exists, like npn and pnp bipolar transistors. Indeed, as illustrated in Fig.11 (a), changing the doping polarities of the substrate and the S/D areas results in a “PMOS” device. The channel now consists of holes and is formed if the gate voltage is below the source potential by one threshold voltage. That is, to turn the device on, $V_{GS} < V_{TH}$, where $V_{TH}$ itself is negative. Following the conventions used for bipolar devices, we draw the PMOS device as in Fig.11 (b), with the source terminal identified by the arrow and placed on top to emphasize its higher potential. The transistor operates in the triode region if the drain voltage is near the source potential, approaching saturation as $V_D$ falls to $V_G - V_{TH} = V_G + V_{TH}$.

Fig.12 (c) illustrates the gate-drain voltages required for each region of operation.
The on-resistance can be controlled by the gate-source voltage. In particular, for \( V_G = V_{TH} \), \( R_{ON} = \infty \), i.e., the device can operate as an electronic switch.

Example 3
A cordless telephone incorporates a single antenna for reception and transmission. Explain how the system must be configured.

Solution
The system is designed such that the phone receives for half of the time and transmits for the other half. Thus, the antenna is alternately connected to the receiver and the transmitter in regular intervals, e.g., every 20 ms (Fig.13). An electronic antenna switch is therefore necessary here. Note: Some cellphones operate in the same manner.
Chapter Summary

- A voltage-dependent current source can form an amplifier along with a load resistor. MOSFETs are electronic devices that can operate as voltage-dependent current sources.
- A MOSFET consists of a conductive plate (the “gate”) atop a semiconductor substrate and two junctions (“source” and “drain”) in the substrate. The gate controls the current flow from the source to the drain. The gate draws nearly zero current because an insulating layer separates it from the substrate.
- As the gate voltage rises, a depletion region is formed in the substrate under the gate area. Beyond a certain gate-source voltage (the “threshold voltage”), mobile carriers are attracted to the oxide-silicon interface and a channel is formed.
- If the drain-source voltage is small, the device operates a voltage-dependent resistor. As the drain voltage rises, the charge density near the drain falls. If the drain voltage reaches one threshold below the gate voltage, the channel ceases to exist near the drain, leading to “pinch-off.”
- MOSFETs enter the “saturation region” if channel pinch-off occurs. As the drain voltage exceeds $V_{GS} - V_{TH}$ and pinch-off occurs
- MOSFETs operating in the saturation region behave as current sources and find wide application in microelectronic circuits.
**Light Emitting Diodes (LEDs)**

LEDs are **p-n junction** devices constructed of gallium arsenide (GaAs), gallium arsenide phosphide (GaAsP), or gallium phosphide (GaP). Silicon and germanium are not suitable because those junctions produce heat and no appreciable IR (IR: Infra Red light) or visible light.

The junction in an LED is **forward biased** and when electrons cross the junction from the n- to the p-type material, the electron-hole recombination process produces some photons in the IR or visible light in a process called **electroluminescence**. An exposed semiconductor surface can then emit light.

**Generation of light**

Photons may be emitted if an electron and hole recombine by a direct band-to-band recombination process in a direct bandgap material. The emission wavelength is:

$$
\lambda = \frac{hc}{E_g} = \frac{1.24}{E_g} \mu m
$$

where, $E_g$ is the bandgap energy measured in electron-volts.

When a voltage is applied across a pn junction, electrons and holes are injected across the space charge region where they become excess minority carriers. These excess minority carriers diffuse into the neutral semiconductor regions where they recombine with majority carriers. If this recombination process is a direct band-to-band process, photons are emitted. The diode diffusion current is directly proportional to the recombination rate, so the output photon intensity will also be proportional to the ideal diode diffusion current.
LED Construction –
Aim – 100% light emitting efficiency

- Important consideration - radiative recombination must take place from the side of the junction nearest to the surface to reduce reabsorption.
- Carrier from n must be injected into the p-side efficiently.
- Consider the fraction of the total diode current that is carried by electrons being injected into the p-side of the junction ($\eta_e$)

Why $n^+\text{-}p$?

\[
\eta_e = \frac{D_e n_p / L_e}{D_e n_p / L_p + D_h p_n / L_h}
\]

\[
\text{Or } \eta_e = \frac{D_h p_n L_e}{D_e n_n L_p}
\]

\[
\eta_\theta = \frac{\mu_e n_h L_p}{1 + \mu_h p_n L_\theta}
\]

If we use Einstein’s equation to substitute this equation. Then,

III-V compounds, $\mu_e \gg \mu_h$

then, $\eta_e$ needs to be close to unity. This can be circumvent by doping n with more electrons ($n \gg p$)

If ($n_n \gg p_n$) = one sided junction $n^+\text{-}p$ diode
LED Characteristic

- The energy of an emitted photon = to the size of the band gap
- BUT this is a simplified statement.
- The energy of an emitted photon from LED is distributed appropriately according to the energy distribution of electrons on the conduction band and holes in the valance band.
- You need to know the distribution of electrons and holes in the CB and VB respectively.

LED construction and energy bandgap:

In order to prevent reabsorption, the upper layer (one that is above the active region) needs to have higher band gap therefore the emitted photons will not be absorbed by the upper layer (photons will be absorbed when $E_p < E_g$).

![LED diagram](image)

**Active region** – Photons will not be absorbed by the n-AlGaAs since the band gap is much higher than GaAs