§9 – Semiconductor Devices

9.1 Semiconductor Devices

We construct semiconductor devices by a process of ***doping***: the intentional addition of impurities.

Certain impurities and imperfections in a semiconductor crystal can **strongly** effect the electrical properties of the semiconductor.

For example, doping silicon in the proportion of 1 part in 10^5 increases the conductivity by a factor of 10^3 at room temperature.

A material is usually considered “ultra-pure” if it contains less that 1 part per million (1ppm) dopant atoms.

If impurities are electrically active, meaning that they contribute or remove electrons then they can override all of the intrinsic effects. Because it was not immediately realised, the formation of a semiconductor industry was delayed.

The most common doping method is to use impurities that can be incorporated into the host lattice.

This means that the doping atoms must be approximately the same size as the host atoms. As such, they are nearby in the periodic table.

In addition, they must have the same chemical configuration, so that they can form covalent bonds.

For example, a phosphorus atom, to a good approximation, acts as a substitute for a silicon atom, without adversely effecting the lattice.

a) Donors

Phosphorus is the next element in the periodic table after silicon:

Si: $3s^2\ 3p^2$ Z=14

P: $3s^2\ 3p^3$ Z=15

Thus, phosphorus has one more nuclear charge and one more electron.

A ***dilute*** material typically has a doping concentration of around 1 part in 10^5 or greater.

An extra electron does not contribute to covalent bonding, but it **does** contribute to the counting of electron states.

Silicon has a valency of 4, meaning that each silicon atom has four valence electrons. Phosphorus (as well as arsenic and antimony) are pentavalent. As such, there is one additional electron, which is given up to the conduction band. Impurities that give up an electron in this way are called ***donors***.

The system stays charge-neutral because the electron remains in the crystal.

The electron concentration in a donor material is generally **much greater** than the intrinsic concentration.

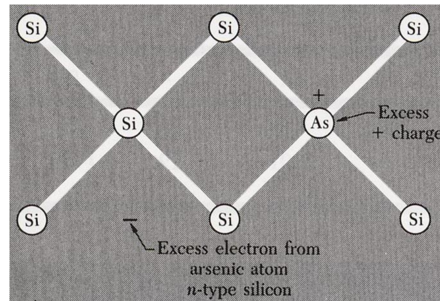
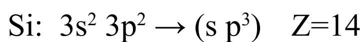


Figure 9.1: Charges associated with an arsenic impurity in silicon. Four of arsenic's five valance electrons form tetrahedral valance bonds similar to silicon, with the final electron available for conduction.

b) Acceptors

Now consider the addition of a trivalent impurity, such as aluminium to a silicon crystal (boron, gallium or indium are other possible trivalent impurities).


Si: $3s^2\ 3p^2 \rightarrow (s\ p^3)$ Z=14

Al: $3s^2\ 3p$ Z=13

This time, there is one less electron than is needed, so an electron is promoted from the valence band, leaving a hole.

Impurities that gain an electron from the valence band are called ***acceptors***.

Doping with high levels of **donor** impurities forms ***n-type*** semiconductors (**n**egative charge carriers).

Doping with high levels of **acceptor** impurities forms ***p-type*** semiconductors (**p**ositive charge carriers).

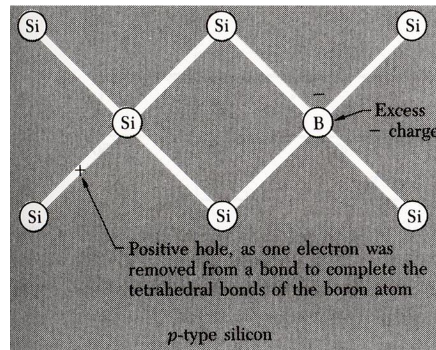


Figure 9.2: Charges associated with a boron impurity in silicon. Since boron has only three valance electrons compared to silicon's four, it ,must complete its tetrahedral bonds by taking an electron from a Si-Si bond. The hole left behind in the silicon valance band is then available for conduction.

9.2 Bound States

a) Bound Donor States

The electron moves in the Coulomb potential $\frac{e}{4\pi\epsilon_0\epsilon_r r}$ of the impurity ion, where for a covalent crystal, ϵ_r is the relative permittivity of the medium, which accounts for the reduction in the Coulomb force between charges due to the polarization of the medium.

This treatment is valid for orbits which are large in comparison to the atomic separation and for orbital frequencies low compared to the band gap frequency.

We can estimate the ionisation energy of the donor impurity using the Bohr model of the hydrogen atom.

In free space, the Bohr model tells us that we can equate the Coulomb force and the centripetal force of the orbits:

$$\frac{e^2}{4\pi\epsilon_0 r^2}=\frac{m v^2}{r} \tag{9.1}$$

Where m is the rest mass of the electron.

This can be used to show that the ionisation energy of atomic hydrogen is given (in SI units) by:

$$E=\frac{e^4 m}{2(4\pi\epsilon_0\hbar)^2}=13.6\,\text{eV} \tag{9.2}$$

Replacing $e^2 \rightarrow \frac{e^2}{\epsilon_r}$ and $m \rightarrow m_e$ into equation (9.2) and choosing the energy to be positive, gives the (approximate) ionisation energy of the donor impurity as:

$$E_d = \frac{e^4 m_e}{2(4\pi\epsilon_r\epsilon_0\hbar)^2} \quad (9.3)$$

The Bohr model also gives the atomic radius of the ground state of hydrogen (the Bohr radius) as:

$$a_0 = \frac{4\pi\epsilon_0\hbar^2}{m e^2} = 0.53\,\text{Å} \quad (9.4)$$

Thus, the atomic radius of the donor is:

$$a_d = \frac{4\pi\epsilon_r\epsilon_0\hbar^2}{m_e e^2} \quad (9.5)$$

Examples of the relative permittivity and relative mass are below:

Table 9.1: Relative permittivity and relative mass ratio for common semiconductors.

Semiconductor	ε_r	m_e/m
Si	11.7	0.19
Ge	15.8	0.08
GaAs	13	0.07

The effect of the relative permittivity and effective mass can increase the atomic radius by a factor of more than 100 and reduce the binding energy by a factor greater than 1000.

The typical radius of an orbit is around 50Å or 5 nm.

P doping in Si has the binding energy reduced (from 13.6eV) to around 45 meV.

As doping in Ge has its binding energy reduced to around 12.7 meV.

b) Bound Acceptor States

We can use the exact same analogy for holes in the valence band as the electrons, except that we replace m by m_h instead of m_e. Doing so gives the ionisation energy and atomic radius of the acceptor impurity as:

$$E_a = \frac{e^4 m_h}{2(4\pi \epsilon_r \epsilon_0 \hbar)^2} \tag{9.6}$$

$$a_a = \frac{4\pi \epsilon_r \epsilon_0 \hbar^2}{m_h e^2} \tag{9.7}$$

B doping in Si reduces the energy to around 45 meV.

Al doping in Si reduces the energy to around 57 meV.

c) Behaviour of Donor and Acceptor States

To see how the donors and acceptors affect the Fermi level, we have to study the band locations:

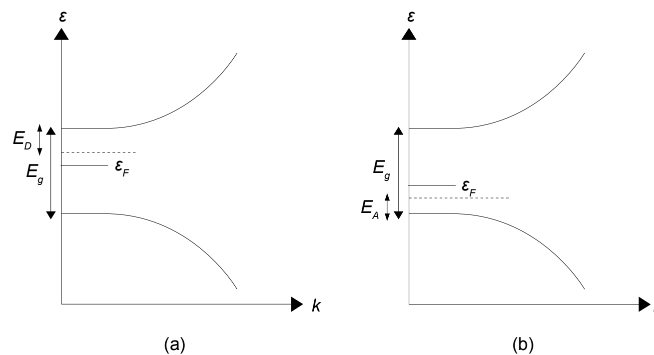


Figure 9.3: Dispersion curves of conduction and valance bands. The dashed lines represent the (a) Bound donor state. (b) Bound acceptor state.

Note:

Donors and acceptors can form a genuine band if the concentration is high enough, especially since the wavefunctions are so large.

At $T=300\,\text{K}$, a significant number of donors/acceptors are ionised to allow pumping of electrons/holes into the other band. This corresponds to $k_B T = 25\,\text{meV}$ and $E_D \sim 40\,\text{meV}$.

The Fermi level is near the small gap between the impurity band and the nearest band.

Remember that the law of mass action can be written as $n\,p=(n_i)^2$. We can also define $\Delta n=n-p$ as the difference in electron and hole concentrations. This will be positive for *n*-type semiconductors and negative for *p*-type. Thus, we can determine the electron and hole concentrations using the following useful relation:

$$\left.\begin{matrix} n \\ p \end{matrix}\right\} = \tfrac{1}{2}\left(\Delta n^2 + 4n_i^2\right)^{\frac{1}{2}} \pm \tfrac{1}{2}\Delta n \qquad (9.8)$$

If we denote the mid-point energy by μ_i, we can write the electron and hole number densities as:

$$n=n_i \exp\left(\frac{(\mu-\mu_i)}{k_B T}\right) \qquad (9.9)$$

$$p=n_i \exp\left(\frac{(\mu_i-\mu)}{k_B T}\right) \qquad (9.10)$$

We can subtract these to give a relation between the concentration difference and the intrinsic concentration:

$$\frac{\Delta n}{n_i}=2\sinh\left(\frac{(\mu-\mu_i)}{k_B T}\right) \qquad (9.11)$$

Again, this can be positive or negative.

Equation (9.11) allows us to see how far from the mid-gap the Fermi level moves when we introduce doping.

For example, for silicon, $n_i=4.5\times10^9\,\mathrm{cm}^{-3}$ and $n_d=\Delta n=10^{16}\,\mathrm{cm}^{-3}$. So, we have $\frac{\Delta n}{2n_i}=10^6$, which gives $\frac{\mu-\mu_i}{k_B T}=14.5$. At room temperature ($T=300\,\mathrm{K}$), this gives the difference between the chemical potential and intrinsic Fermi level $\mu-\mu_i=0.36\,\mathrm{eV}$.

Since the band gap is 1.1 eV, the Fermi level moves considerably towards the donor level.

There is a crossover from intrinsic to extrinsic behaviour seen when we vary temperature.

Let us consider *n* (we will have a similar result for *p*).
Assume that Δn has been fixed by doping.

Remember that the intrinsic concentration is given by $n_i = 2\left(\frac{k_B T}{2\pi\hbar^2}\right)^{\frac{3}{2}} (m_e m_h)^{3/4} \exp\left(-\frac{E_g}{2 k_B T}\right)$.

Substituting into equation (9.9), we can plot $\ln n$ against the inverse of $k_B T$ and consider the form of the intrinsic case:

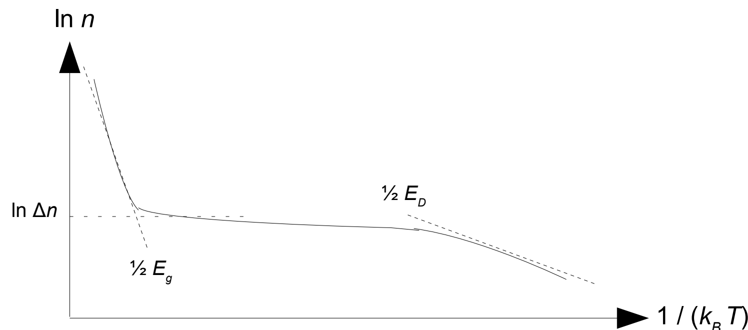


Figure 9.4: Graph showing the natural logarithm of electron concentration against $1/k_B T$ in a semiconductor with donor impurities. At high temperatures (left of graph), intrinsic behaviour occurs. The crossover between intrinsic and extrinsic behaviour occurs at ln(n)=ln(Δ*n*). At very low temperatures, there is a freeze out of electrons onto donors.

The crossover between intrinsic and extrinsic behaviour is mainly seen in the conductivity.

9.3 Conductivity

We saw previously that the electrical conductivity of a metal depends on the carrier density:

$$\sigma = \frac{n e^2 \tau}{m} \tag{9.12}$$

This relation works for semiconductors too, but we usually prefer to separate contributions from electrons and holes.

$$\sigma = n e\left(\frac{e\tau_e}{m_e}\right) + p e\left(\frac{e\tau_h}{m_h}\right) = n e \mu_e + p e \mu_h \tag{9.13}$$

Where μ_e and μ_h are the electron and hole ***mobility***. The mobility is usually expressed in CGS units as $cm^2\ V^{-1}\ s^{-1}$.

The mobility is defined as the drift velocity per unit electric field:

$$\mu = \frac{v_g}{E} \tag{9.14}$$

Where v_g is the drift velocity: the average velocity that a carrier attains due to an electric field.

As with resistivity, we can split the mobility (or rather its reciprocal) into contributions from scattering from impurities and scattering from phonons in the crystal:

$$\frac{1}{\mu} = \frac{1}{\mu_L} + \frac{1}{\mu_i} \tag{9.15}$$

The mobility is mainly determined by the lifetime.

The purest materials tend to have the highest mobility. Below are examples of electron mobilities in pure semiconductors:

Table 9.2: Electron mobility of pure semiconductors.

Semiconductor	Electron Mobility ($cm^2/V\ s$)*
Si	1350
Ge	3600
GaAs	8000
PbTe	2500 **

* The electron mobilities shown above are determined at room temperature.

** The highest mobility observed in a standard semiconductor material is 5×10^6 cm/Vs in PbTe at 4K (Kittel 7th Ed. 1996).

The record mobility is 10^7 in the 2-dimensional electron gas of a Ga As heterostructure.

If $\mu_e \approx \mu_h$, then $\sigma = (n+p) e \mu_e$, which gives an easy way to estimate n.

9.4 The p-n Junction

A p-n junction is constructed from a single crystal modified in two separate regions. Acceptor impurity atoms are incorporated into one section of the crystal to produce the *p* region, in which the majority carriers are holes. Donor impurity atoms are added to the other to produce the *n* region, in which the majority carriers are electrons. The interface region may be as little as 10^{-4} cm thick.

Away from the junction on the *p* side, there are negatively ionised acceptor atoms and an equal concentration of holes. Conversely, on the *n* side, there are positively charged donor atoms and an equal concentration of free electrons.

Holes concentrated on the *p* side would like to diffuse to fill the crystal uniformly and electrons would like to diffuse on the *n* side. But, diffusion will upset the local electrical neutrality of the system.

A small charge transfer by diffusion leaves an excess of ionised acceptors in the *p* region and an excess of ionised donors in the *n* region. This creates an electric field $\vec{E}$ directed from *n* to *p*, with an associated electrostatic potential φ, which prevents further diffusion and maintains separation of the two carrier types. This separation is known as the ***depletion zone***.

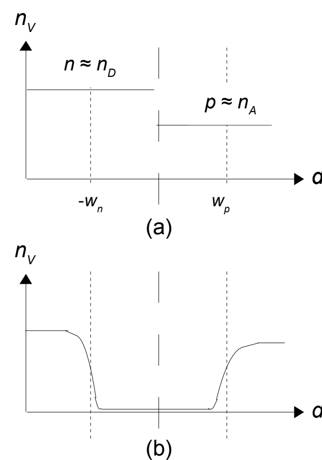


Figure 9.5: Graph of number density versus distance for a p-n junction (a) immediately after formation (b) after formation of depletion zone.

The divergence of the electric field is given by Poisson's equation:

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_r \epsilon_0} \tag{9.16}$$

We also know that the electric field is equal to the negative gradient of the electrostatic potential (if there is no magnetic field):

$$\vec{E} = -\vec{\nabla} \phi \tag{9.17}$$

To keep the equations simple we will model the boundaries as **sharp**, although in reality, there will be a partial ionisation of the dopants near the sides of the depletion zone.

We will also ignore intrinsic effects, assuming that $n_i \ll n_D$ and n_A, although there will be a small carrier concentration in the depletion zone.

We can use electrostatics to determine the depletion widths w_n and w_p of the *n*- and *p*-sections.

Combing equations (9.16) and (9.17) gives:

$$\vec{\nabla}\cdot(-\vec{\nabla}\phi)=-\nabla^2\phi=\frac{\rho}{\epsilon_r\epsilon_0} \tag{9.18}$$

In 1-dimension, the Laplacian becomes $\nabla^2\rightarrow\frac{d^2}{d\,x^2}$. We also know that the charge density *ρ* is given by $\rho=n\,q$, where *n* is the number density (n_A or n_D) and *q* is the carrier charge ($\pm e$). Thus, in the *p*-*n* material, the potential must satisfy:

$$\frac{d^2}{dx^2}\phi(x)=\begin{cases}-\frac{n_D e}{\varepsilon_r\varepsilon_0} & -w_n < x \le 0\\ +\frac{n_A e}{\varepsilon_r\varepsilon_0} & 0\le x < w_p\\ 0 & \text{elsewhere}\end{cases} \tag{9.19}$$

Assume the potential will be of the form:

$$\phi(x)=A+\frac{1}{2}\frac{n\,q}{\epsilon_r\epsilon_0}(x+B)^2 \tag{9.20}$$

Any expression of this form is a valid solution to equation (9.19) within the limits of the integral.

We want $\phi(-w_n)=\phi'(-w_n)=0$, so set $A=0,\ B=w_n$ into equation (9.20) for the *n*-section:

$$\phi(x)=-\frac{1}{2}\frac{n_D e}{\epsilon_r\epsilon_0}(x+w_n)^2;\quad -w_n\le x\le 0 \tag{9.21}$$

For the *p*-section, we want $\phi'(w_p)=0$, and we also need to include the potential drop $\Delta\phi$, so set $A=-\Delta\phi,\ B=-w_p$ into equation (9.20):

(9.22)

$$\phi(x) = -\Delta\phi + \frac{1}{2}\frac{n_A e}{\epsilon_r \epsilon_0}(x - w_p)^2; \quad 0 \le x \le w_p$$

Now, equations (9.21) and (9.22) are equal at $x = 0$:

$$\phi(0) = -\frac{n_D e}{\epsilon_r \epsilon_0}\frac{w_n^2}{2} = -\Delta\phi + \frac{n_A e}{\epsilon_r \epsilon_0}\frac{w_p^2}{2}$$

Rearranging gives the potential drop in terms of the depletion widths:

$$\Delta\phi = \frac{e}{2\epsilon_r \epsilon_0}\left(n_D w_n^2 + n_A w_p^2\right) \tag{9.23}$$

We can also equate the gradients of equations (9.21) and (9.22) evaluated at $x = 0$:

$$-\frac{n_D e}{\epsilon_r \epsilon_0} w_n = -\frac{n_A e}{\epsilon_r \epsilon_0} w_p$$

$$n_D w_n = n_A w_p \tag{9.24}$$

This is the same condition that the surface areas be equal under $\rho(x)$.

Substituting w_p from equation (9.24) into equation (9.23) gives:

$$\Delta\phi = \frac{e}{2\epsilon_r \epsilon_0}\left(n_D w_n^2 + n_A \left(\frac{n_D}{n_A}\right)^2 w_n^2\right) = \frac{e}{2\epsilon_r \epsilon_0}\frac{n_D(n_A + n_D)}{n_A} w_n^2$$

Rearranging gives the depletion width in the n-region:

$$w_n = \left(\frac{2\epsilon_r \epsilon_0 n_A \Delta\phi}{e\, n_D (n_A + n_D)}\right)^{\frac{1}{2}} \tag{9.25}$$

Similarly, substituting w_n from equation (9.24) into equation (9.23) gives:

$$\Delta\phi = \frac{e}{2\epsilon_r \epsilon_0}\left(n_D \left(\frac{n_A}{n_D}\right)^2 w_p^2 + n_A w_n^2\right) = \frac{e}{2\epsilon_r \epsilon_0}\frac{n_A(n_A + n_D)}{n_D} w_p^2$$

$$w_p = \left(\frac{2 \epsilon_r \epsilon_0 n_D \Delta \phi}{e n_A (n_A + n_D)} \right)^{\frac{1}{2}} \tag{9.26}$$

So, there is a direct coupling between the potential drop across the junction and the widths of the depletion zone on the two sides of the junction.

9.5 Zero Voltage Bias

If there are no **external** potential differences placed across the junction, then the potential difference across the junction is set by the change in the chemical potential caused by the doping.

We can consider two analogous pictures:

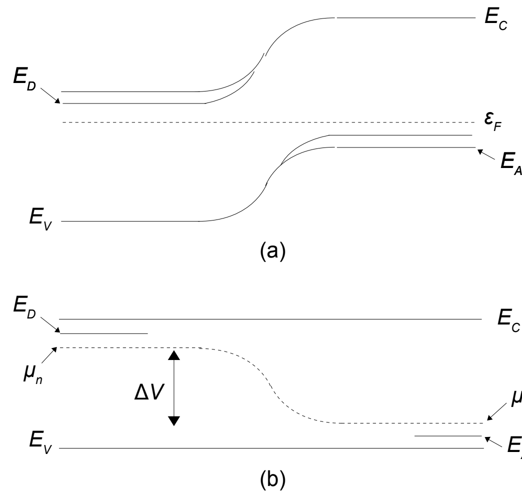


Figure 9.6: Analogous representations of zero voltage bias in a p-n junction. (a) The Fermi level is taken as constant by varying the position of the conduction and valance bands. (b) Conduction and valance bands are taken as constant by varying the Fermi level. In this case, the potential difference (for energy in electron volts) is the difference in chemical potential of the *n* and *p* sections.

Provided that $(n_D, n_A) \gg n_i$, we can take the two limits of the hyperbolic sine functions defined by equation (9.11):

$$n_D = n_i \exp\left(\frac{(\mu_n - \mu_i)}{k_B T} \right) \quad n\text{-type} \tag{9.27}$$

$$n_A = n_i \exp\left(\frac{(\mu_i - \mu_p)}{k_B T} \right) \quad p\text{-type} \tag{9.28}$$

Multiply equation (9.27) by (9.28), noting that $\Delta\phi = \frac{(\mu_n - \mu_p)}{e}$:

$$n_D n_A = n_i^2 \exp\left(\frac{(\mu_n - \mu_p)}{k_B T}\right) = n_i^2 \exp\left(\frac{e\Delta\phi}{k_B T}\right) \tag{9.29}$$

Taking logarithms to base *e* on both sides and rearranging, gives the potential drop as:

$$\Delta\phi = \frac{k_B T}{e} \ln\left(\frac{n_A n_D}{n_i^2}\right) \tag{9.30}$$

This allows us to calculate $\Delta\phi$, which in turn, allows us to determine the depletion widths w_n and w_p .

Example in Si at $T = 300\text{K}$, $n_i = 2.1\times10^{19}\,\text{cm}^{-3}$ and $k_B T = 25\,\text{meV}$. Let $n_A = n_D = 10^6\,\text{cm}^{-3}$. Then, $\Delta\phi = 0.73\,\text{eV}$.

Furthermore, we know that in SI units, $\epsilon_0 = 8.9\times10^{-12}$, $e = 1.6\times10^{-19}\,\text{C}$ and for Si, $\epsilon_r = 11.7$. So, the depletion widths are $w_n = w_p = 2.2\times10^{-7}\,\text{m} = 220\,\text{nm}$.

Typical depletion widths are around this size.

If we increase the doping in the above example to 1 part in 10^5, then the potential drop hardly changes, since the depletion widths decrease as a square root.

These are decent numbers for engineering purposes, because we need to be able to construct *p-n* devices with junctions that are well defined in this region (around room temperature).

9.6 Applied Voltage Bias

Let us place a *p-n* junction in a circuit with a chemical battery and apply a voltage φ to the *p*-section.

The voltage drop appears ***entirely*** at the depletion zone because both doped regions conduct well, whereas the depletion zone is insulating.

The first thing that happens when the field is applied is that the depletion zone width changes in response to the applied potential:

(9.31)

$$w_n = \left(\frac{2\epsilon_r \epsilon_0 n_A (\Delta\phi - \phi)}{e n_D (n_A + n_D)} \right)^{\frac{1}{2}}$$

For **forward** bias, $\phi > 0$, so the depletion zone gets **smaller**.

For **reverse** bias, $\phi < 0$, so the depletion zone gets **larger**.

This turns out to be relatively unimportant, unless the depletion zone touches the ends of the device. In this case, reverse breakdown would result.

The key to understanding the *p*-*n* junction is to examine the current.

9.7 Zero-Bias Current

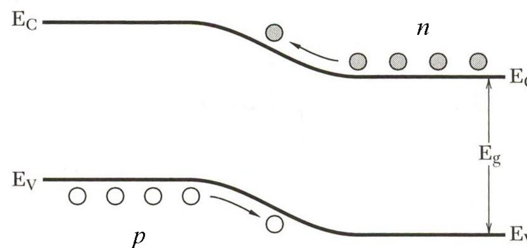


Figure 9.7: Illustration of electron and hole zero bias current in a *p*-*n* junction. The shaded circles indicate electrons and the unshaded circles holes. The arrows indicate the direction of forward current. For reverse current, the arrows would be reversed.

There are several names for the two currents (densities):

j_F	j_R
Forward	**Reverse**
Uphill	Downhill
Diffusion	Field/Drift
Recombination	Generation

The last two names are perhaps the most descriptive.

By definition, at zero bias, the system is in thermal equilibrium, so we have:

$$j_F = j_R \tag{9.32}$$

The recombination current is caused by carriers in outlying regions of the Fermi-Dirac distribution gaining enough energy to climb the potential "hill". There are many carriers because of the doping but there are generally few with $\epsilon > (\Delta\phi - \phi)$. The exact number is dependant on φ. It is called the *recombination* current because when the electrons reach the *p*-section (or holes reach the *n* -section), they recombine with a hole (or electron), causing spontaneous emission of a photon. The *generation* current is so-called because it is formed by spontaneous generation of ***minority*** carriers on the opposite side of the junction. Once they have been created, they are immediately attracted by the field to the other side of the junction.

This process only depends on the band gap E_g, i.e. does not depend on φ or $\Delta\phi$.

9.8 Applied Current Bias

As we previously suggested, the forward current varies with applied field, whereas the reverse current does not.

a) Reverse Current

Let us first estimate the generation current. We will make two key assumptions:

1. Only carriers within one diffusion length from the edge of the depletion zone contribute, i.e. $x < L$;

   Inside the depletion zone, there are more carriers, but $w \ll L$.

   Outside the depletion zone $x \geq L$ we will assume that scattering occurs before the edge.

2. The generation rate is given by $\frac{n_{min.}}{\tau}$.

   Where: $n_{min.}$ is the minority concentration and τ is the recombination lifetime before an electron finds a hole and recombines.

This gives the reverse current density:

$$j_R = \frac{e\, l\, n_{\min.}}{\tau} \tag{9.33}$$

Where: l is the mean free path.

For electrons in the *p*-section, the minority carrier concentration can be written as:

$$n_{\text{min.}} = \frac{n_i^2}{n_A} \quad (9.34)$$

This is usually small $n_{\text{min.}} \sim 10^4 \,\text{cm}^{-3}$.

In equation (9.33) the ratio $v_F = \frac{l}{\tau}$ is the velocity, which is approximately the sound velocity $v_F \sim 10^4 \,\text{m}\,\text{s}^{-1}$. Using these values the generation current density is around $j_R \sim 10^{-7} \,\text{A m}^{-2}$.

The forward current varies exponentially:

$$j_F = A \exp\left(\frac{e\phi}{k_B T}\right) \quad (9.35)$$

At $\phi = 0$ $j_F = A = j_R$, so we can write the total current density as $j = (j_F - j_R) = j_R \left[\exp\left(\frac{e\phi}{k_B T}\right) - 1\right]$. If we consider the total current density of electrons **and** holes, the total forward current is given by:

$$j = \left(j_{R_e} + j_{R_h}\right)\left(e^{\frac{e\phi}{k_B T}} - 1\right) = j_s\left(e^{\frac{e\phi}{k_B T}} - 1\right) \quad (9.36)$$

9.9 Diode Characteristics

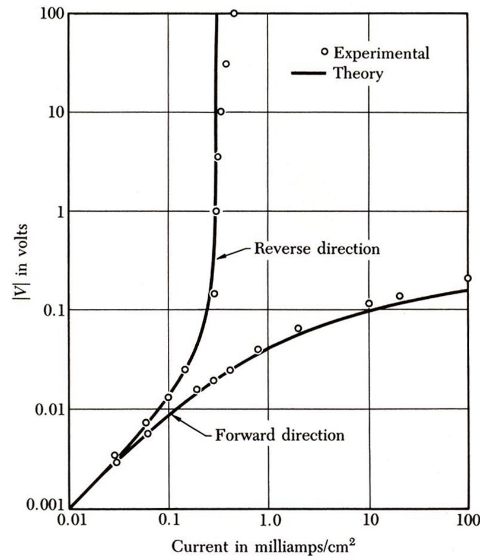


Figure 9.8: Graph of voltage versus current for a diode in forward bias and reverse bias modes. Experimental data points have be plotted along with the experimental curves.

The only parameter we are concerned with here is the reverse-bias current. Let j_R be the theoretical current density and J_R be the practical measurement, which will be proportional to the area of the device.

A diode is an example of a rectifier circuit: current only flows in one direction (although in reality there is a small reverse current).

When we use a diode in a circuit, we can assume the reverse current is zero, and assume that we can approximate the graph of the reverse current in Fig. 9.8 to a Dirac delta function at some diode-drop voltage.

A useful device needs a reverse current density as small as $J_R \sim 10^{-13}\,\text{A mm}^{-2}$, which is unmeasurable.

So, to produce a forward current density of 10 mA mm^{-2} at room temperature, we would need:

$$\phi = \frac{k_B T}{e} = \ln\left(\frac{10^{-2}}{10^{-13}}\right) = 0.63\ \text{V}$$

But, to increase the forward current density by a factor of 10, would only need a voltage of 0.69 V.

So, in general, we assume that V_D is constant.

9.10 Field Effect Transistor (FET)

We now know that a *p-n* junction has two important characteristics:

1. A variable width insulating depletion zone;

2. A exponential relation between forward current and voltage.

In the 1950s, there was great interest in solid state physics to produce the first transistor: a three-terminal device that could produce "gain". In other words, this is a semiconductor device that is able to take a low power signal on input and amplify the signal to produce a higher power signal on output.

The original proposal made by Shockley was a ***field effect transistor*** (***FET***), but there were problems getting the device to work. The first viable working transistor was a ***bipolar junction transistor*** (***BJT***), which will be considered later.

The simplest type of field effect transistor is the ***junction field effect transistor*** (***JFET***), which is constructed from a *p-n* junction.

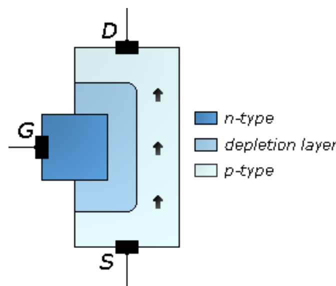


Figure 9.9: A p-type Junction FET (JFET). The device is labelled p-type or n-type based on the bulk material that the gate is constructed from. For p-type JFETs, the gate is produced from n-type material. The current flow consists of the majority carriers.

The JFET is constructed by connecting the bulk section or ***substrate*** (either *n* or *p*) to a circuit. The input connection is known as the ***source*** and the output known as the ***drain***. We then add another connection to the second section of the junction. This is known as the ***gate***. If we apply a voltage V_{DS} across the channel, then for an *n*-type JFET, electrons move from the (−ve) source to the (+ve) drain, whereas for a *p*-type JFET, the holes move from the (+ve) source to the (−ve) drain.

If we make the gate reverse biased with respect to the channel, then the depletion zone will increase its size and the current will diminish for a fixed voltage applied across the channel. If we make the reverse bias large enough, the depletion zone gets so large that it *pinches-off* the drain current.

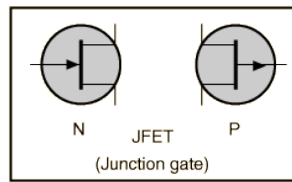


Figure 9.10: Circuit convention for n-type and p-type JFETs.

The drain current I_D can be found by:

$$I_D = I_{DSS}\left(1 - \frac{V_{GS}}{V_p}\right)^2 \tag{9.37}$$

Where: V_p is the pinch-off voltage, V_{GS} is the gate-to-source voltage and I_{DSS} is the drain current with the gate shorted to source (i.e. $V_{GS}=0$).

9.11 Bipolar Junction Transistor (BJT)

Bipolar junction transistors were the main type of transistors used in the semiconductor industry until around the 1970s, when ***complementary metal-oxide semiconductor*** (CMOS) technology took off. CMOS originally used ***metal oxide semiconductor field effect transistors*** (MOSFETs) and allowed low power, high density circuits not possible with BJTs. CMOS is now the predominant technology in digital integrated circuits such as microprocessors. However, there are still many instances in which BJTs are still preferable e.g. radio frequency circuits in wireless systems.

Instead of a source, drain and gate, the three connections in the BJT are called the ***base***, ***emitter*** and ***collector***.

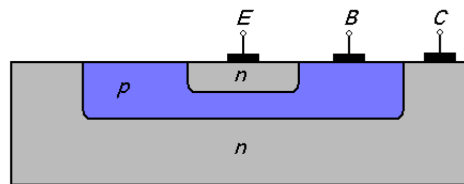


Figure 9.11: Simplified cross-section of an *npn* (bipolar junction) transistor. The emitter and collector regions are n-doped, whereas the base is p-doped. In a *pnp* transistor, the doping is reserved.

In typical operation, the base-emitter junction is **forward** biased and the base-collector junction is **reverse** biased.

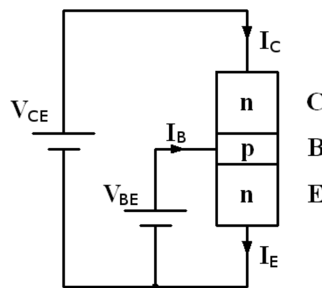


Figure 9.12: Simplified circuit diagram of an npn transistor circuit. The arrows indicate conventional current system

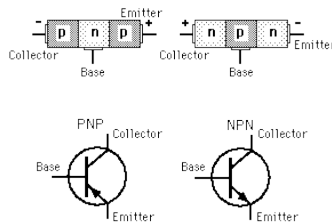


Figure 9.13: Circuit convention of pnp and npn transistors and pictorial representation of how they are connected in a circuit.

When a voltage higher than the ***cut-in voltage*** (around 600-700 mV for Si) is passed between the base and emitter, the base-emitter junction is "turned on", allowing charge pumping of majority carriers (electrons in the *npn* case) from the emitter into the base. The base region is formed of lightly doped, high resistivity material, whereas the emitter region is heavily doped. This means that the total emitter current is **much** higher than from the base to the emitter.

The electric field between the base and collector (caused by V_{CE}) causes minority carriers (electrons in the *npn* case) to be transferred to the base to cross the base-collector junction, forming a collector current I_C. Those carriers left, exit the base terminal to form the base current I_B (along with any majority carriers injected into the emitter).

The base is made long and thin, so the minority carriers spend little time in the base, and few have time to recombine with majority carriers.
The ratio of the collector and emitter currents is denoted by:

$$\alpha = \frac{I_C}{I_E} \tag{9.38}$$

This should be very close to unity for an efficient BJT.

More importantly, the ***current gain*** is the ratio of the collector-emitter to base-emitter currents:

$$\beta = \frac{I_C}{I_B} \tag{9.39}$$

This is set by the geometry of the BJT and is typically equal to100 (so an input current of 1 mA leads to an output current of 100 mA).

The emitter current is simply the sum of the base and collector currents:

$$I_E = I_B + I_C \tag{9.40}$$

9.12 Solar Cells and Light-Emitting Diodes

Photons are involved in any inter-band process:

$$\text{electron} + \text{hole} \rightarrow \text{photon} \tag{9.41}$$

(This suggests a hole is like an imaginary positron)

Normally, there is a thermal equilibrium with the electromagnetic field. However, we can attach an external circuit to tip the balance one way or the other, causing absorption or emission of photons. *Light-emitting diodes* (LEDs) are semiconductor devices that emit incoherent photons(unlike lasers) when a forward voltage bias is applied. This is due to the ***electroluminescent effect***.

The colour of the light depends on the chemical properties of the semiconductor being used.

A ***photovoltaic cell***, commonly known as a ***solar cell***, is a semiconductor device that is able to use photons absorbed at a certain wavelengths (usually in the visible part of the spectrum) to excite charge carriers to produce a voltage in the same way as a usual diode. This is known as the ***photovoltaic effect***.
We prefer to use direct-gap semiconductors to produce LEDs or solar cells, otherwise we need to wait for the phonons to couple.

Semiconductors with wider band gaps than Si or Ge couple to visible light.

For example, GaAs (1.4 eV), AlAs (2.2 eV) or an alloy of the two.

The lifetime of the carriers, once generated, are quite long, $\tau \sim 10^{-7}\,\mathrm{s}$.

The diffusion length before recombination is also long $l \sim 10\,\mu\mathrm{m} \gg w$, where w is the depletion

zone width.

Thus, the carriers have ample opportunity to cross the depletion zone, if the field allows them.

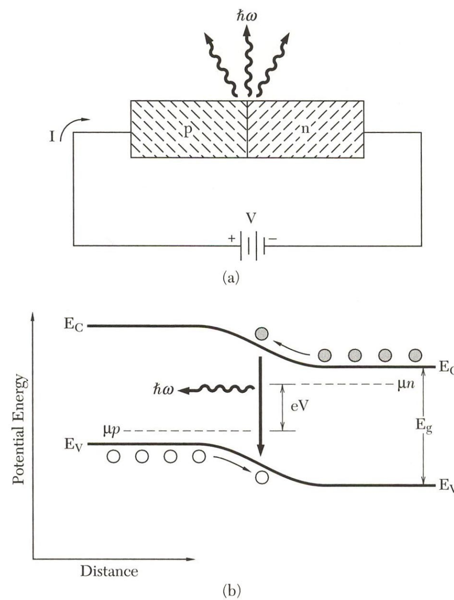


Figure 9.14: Electron-hole recombination into photons across a p-n junction. If the material is a direct gap semiconductor, this could represent a LED. (b) could also represent a photovoltaic cell if the direction of all the arrows are reversed (including that of the photon).

9.13 Heterostructure

Semiconductor ***heterostructures*** are layers of two or more different semiconductors grown coherently, with one common crystal structure. If there are just two layers (one interface) then the material is known as a ***heterojunction***.

A heterostructure may be seen as a single crystal in which the occupancy of the atomic sites change at the interface. For example, one side may be *Ge* and the other *GaAs*. Both have the same atomic spacing but *Ge* has a diamond structure, whereas *GaAs* has a zincblende structure. However, both form tetrahedral bonds, so they fit together coherently, as if they were a single crystal.

The band gaps of *Ge* and *GaAs* are however, very different (namely 0.67 and 1.43 eV respectively).This constitutes a *normal* band edge offset in Fig. 9.16.

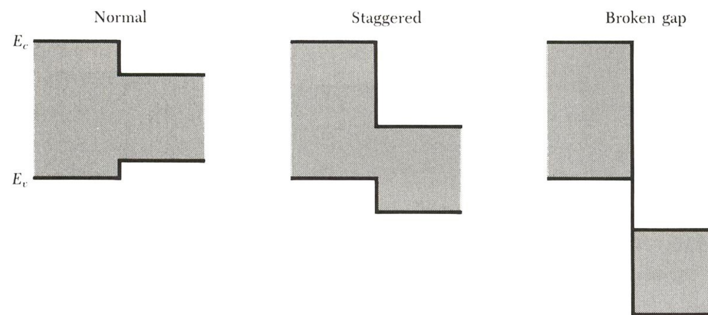


Figure 9.16: Three types of band edge offset at the interface of a heterostructure.

Band edge offsets act as potential barriers in the opposite sense to electron and holes. For a normal offset, the electrons and holes are pushed by the barrier from the wide-gap to narrow-gap region.

a) *n*-*N* Heterojunction

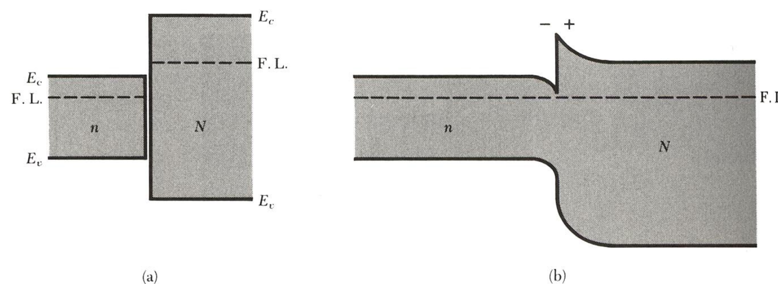


Figure 9.17: (a) Two *n*-type semiconductors with band gaps $N > n$ that are not in diffusive contact. (b) The two semiconductors are now in diffusion equilibrium in a heterojunction. In order for this to be possible, the Fermi level must be independent of position. This is achieved by transferring electrons from the high concentration (N) region to the low concentration (n) region, forming a depletion layer of positively charged ions.

Fig. 9.17 is an example of an *n*-*N* heterojunction. The bending down on the *n*-side and bending up on the *N*-side is a result of the electron transfer. This forms a potential well for electrons. It is possible for the potential well to dip below the Fermi level. In this case, electrons in the well are restricted from moving in the direction normal to the interface, so they can only move along the plane parallel to the heterojunction interface. This is known as a ***two-dimensional electron gas*** (***2DEG***).

If the doping of the *n*-side is reduced to a negligible value, there will be very few ionised donors of that side of the electron-rich layer. The mobility of such electrons is largely limited by lattice scattering, which, as mentioned previously, drops off sharply with temperature. We can use this to

create high mobility electrons at low temperatures.

b) Semiconductor Laser

Stimulated emission of radiation can occur in direct-gap semiconductors from the radiation emitted when electrons spontaneously recombine with holes. For the rate of stimulated (coherent) emission to exceed the spontaneous rate an "inverted population" of electron and hole concentrations must be created by the excitation current. The recombination times for the excess carriers are much longer than the times for the conduction electrons and holes to reach thermal equilibrium in the conduction and valance bands.

This steady state condition for electron and hole populations is described by separate Fermi levels μ_c, μ_v, which are referred to as ***quasi-Fermi levels***.

For population inversion, we require that:

$$\mu_c > \mu_v + \epsilon_g \tag{9.42}$$

Virtually all practical injection lasers employ a double heterostructure. Here, the lasing semiconductor is embedded between two wider-gap semiconductor regions with opposite doping. A common example is *GaAs* embedded in (*Al*,*Ga*)*As*, as shown in Fig.9.18. In such a structure, there is a potential barrier that prevents the outflow of electrons to the *p*-type region and an opposite potential barrier that prevents the outflow of holes to the *n*-type region.

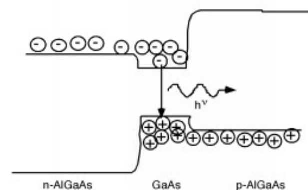


Figure 9.18: Double heterostructure injection layer of a semiconductor laser.