ELEC1205: SOLID STATE DEVICES

Outlined in these notes is an overview of the information required for ELEC1205; this is not everything you need! The lectures will provide more detail, especially for derivation of equations. The text books for the module are (1) Solid State Electronic Devices, B.G. Streetman and S.K. Banerjee, (2) Introductory Semiconductor Device Physics, G. Parker – it is advisable to consult these. You may find, Neamen, Semiconductor Physics and Devices, McGraw-Hill, 3rd Edition onwards, useful also.

INTRODUCTION

There are 200 “common semiconductors” and more than 2000 inorganic semiconductors in total; these complex materials have many interesting properties. Semiconductor devices often demand very specific properties that only certain semiconductors can provide. Silicon is easily the most commonly used semiconductor (95% of all devices) and is used for a wide range of electronic devices; GaAs and GaN are used for light-emitting device because they are good at emitting light, CuInSe₂ is a useful material for solar cells because it is good at absorbing light. Each of the 200 common semiconductors has its own particular use and many are the basis of very active research.

In addition to inorganic semiconductors there are emerging fields of organic semiconductors and carbon nanotubes. These materials bring many exciting new ideas, the story of semiconductor devices is about 100 years old, but the story is far from over. The development of semiconductor devices will continue to drive consumer, medical and military electronics over many years to come.

In this module we seek to understand the fundamental properties of semiconductors, we need to understand the physics that makes semiconductor devices operate and we will look at the basic operation of semiconductor devices. To begin we need to review our understanding of the structure of the atom. In doing so we consider how the structure of the atom has been established historically by others. I have included the bare basics below:

Planck observed black-body radiation, is emitted as discrete units of energy (termed quanta where the energy units are $h\nu$) from a heated metal sample. (Where $h$ is Planck’s constant (6.63 x $10^{-34}$ Js) and $\nu$ is frequency of emitted radiation.)

Einstein then interpreted the experiment demonstrating discrete quantization of light where a metal was illuminated by monochromatic light. Some electrons in the metal (under vacuum) receive sufficiently energy to be ejected – this is termed the photoelectric effect. The maximum energy of the ejected electron ($E_m$) is related to the frequency ($\nu$) of the incident light as follows:

$$E_m = h\nu - q\Phi$$

$E_m$ = maximum energy for emitted electrons, $\nu$ = frequency of the radiation. $q$ = magnitude of the electronic charge, $\Phi$= (in volts) is characteristic of the metal used, $q\Phi$ is the work function and represents the minimum energy for an electron to escape from the metal.

The quantized units of light energy were considered as localized packets of energy, called photons. Light is shown to have both a wave and a discrete nature – this wave-particle duality is fundamental to quantum processes.

Based upon this duality de Broglie proposed that particles (such as electrons) could manifest a wave character in certain experiments and was confirmed by Davisson and Germer who demonstrated
this from the diffraction of electrons by a periodic array of atoms in a crystal. de Broglie stated that a particle of momentum $p=mv$ has a wavelength of:

$$\lambda = \frac{h}{p} = \frac{h}{mv}$$

**Bohr Model**

1. Electrons exist in stable, circular orbits
2. The electron may shift to a higher or lower energy orbit
3. The angular momentum of the electron in an orbit is always an integral multiple of Planck’s constant

$$r_n = \frac{Kn^2\hbar^2}{mq^2}$$

where $r_n = \text{radius of nth orbit}$, $K=4\pi\epsilon_0$, $n=\text{integer of orbit}$, $m=\text{mass of electron}$, $\hbar=h/2\pi$, $\epsilon_0=\text{permittivity in vacuum}$

Total energy of the electron in the nth orbit (note sign)

$$E_n = -\frac{mq^4}{2K^2n^2\hbar^2}$$

Energy difference between orbits (shown here for Lyman series)

$$E_{n2} - E_{n1} = \frac{mq^4}{2K^2\hbar^2} \left( \frac{1}{n_2^2} - \frac{1}{n_1^2} \right)$$

Frequency of light ($\nu$) given off by transition between these orbits

$$\nu_{21} = \frac{mq^4}{2K^2\hbar^2} \left( \frac{1}{n_1^2} - \frac{1}{n_2^2} \right)$$

$$\left[ \frac{mq^4}{2K^2\hbar^2} \right] \text{ is the Rydberg constant (R) \times speed of light (c)}$$

Electron transitions and their resulting wavelengths for hydrogen (figure from Wikipedia)
**Heisenberg uncertainty principle**

Since it is impossible to describe with absolute precision events involving individual particles on the atomic scale, any measurement of the position \( (x) \) and momentum \( (p) \) of a particle, the uncertainties in the two measured quantities are related by

\[
(\Delta x)(\Delta p) \geq \frac{\hbar}{2}
\]

And uncertainties in the energy \( (E) \) measurement will be related to uncertainty in time \( (t) \):

\[
(\Delta E)(\Delta t) \geq \frac{\hbar}{2}
\]

The implication of this theory is that we can only obtain the probability of finding an electron at a certain position and because Planck’s constant is small \((6.63 \times 10^{-34} \text{ Js})\) we cannot accurately determine the energy of an electron.

**Schrödinger Wave Equation**

(For free particle in one dimension (which can be generalised to three dimensions))

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

Where \( \Psi \) is the wave function, \( V \) is the potential energy. This time-independent version of the equation can be solved giving a wave that describes the quantum aspects of the system. In essence the wave function \( \Psi_n \) (Eigen function) corresponds to a particle of energy \( E_n \).

Schrödinger’s equation can be solved for hydrogen, but it not easy to solve for other atoms, so we generalise and assume a particle in an infinite box. The particle is assumed to be trapped in a potential well where \( V(x) \) is zero except at the boundaries where \( x=0 \) and \( L \) (infinitely large) so Schrödinger’s equation becomes:

\[
\frac{d^2 \Psi(x)}{dx^2} + \frac{2m}{\hbar^2} E\Psi(x) = 0 \quad \text{where} \quad 0 < x < L
\]

A solution taking into account the boundary condition is

\[
\Psi = A \sin kx, \quad k = \frac{\sqrt{2mE}}{\hbar}
\]

because \( \sin kx \) is zero at \( x=L \). \( A \) is the amplitude of the wave function.

If \( \Psi \) is to be zero at \( x=L \), then \( k \) must be some integral multiple of \( \pi/L \)

\[
k = \frac{\sqrt{2mE_n}}{\hbar} = \frac{n\pi}{L}
\]

Thus for each allowable value of \( n \) the particle energy \( (E_n) \)
Thus the energy is quantized – only certain values of energy are allowed. The integer \( n \) is called the quantum number. The wave function \( \Psi_n \) and the corresponding \( E_n \) describe the quantum state of the particle.

\[
E_n = \frac{n^2 \pi^2 \hbar^2}{2ml^2}
\]

\[
\Psi_n = \sqrt{\frac{2}{L}} \sin \left( \frac{n\pi}{L} x \right)
\]

THE PERIODIC TABLE

The periodic table is an excellent tool when dealing with semiconductors as it can be used to predict many material properties.


Semiconductors are usually produced through covalently bonded combinations of group IV, III-V or II-VI elements. Group I-VII compounds form ionic bonds and hence don’t form semiconductors. Although semiconductors are covalent there are varying degrees of ionicity and only elemental semiconductors (Si, Ge or C) are truly covalent.
It is useful to consider some of the more common semiconductors (Si, Ge, C, GaAs, GaN, InP, GaP, ZnSe, CdTe, CdS, ZnO) and look at their positions in the periodic table, patterns begin to emerge that allow us to make predictions with regard to semiconductor properties.

It is perhaps possible to appreciate that GaAs and GaP might have similar properties, as might CdSe and ZnSe, equally there are similarities between GaN and ZnO.

Now if we consider the **Pauli Exclusion Principle** in the context of the Periodic Table, from the **Pauli Exclusion Principle** we know that *no two electrons in an interacting system can have the same set of quantum numbers.*

3 quantum numbers $n, l, m$

<table>
<thead>
<tr>
<th>$n$</th>
<th>$l$</th>
<th>$m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td>2</td>
<td>-2</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>$\pm \frac{1}{2}$</td>
</tr>
</tbody>
</table>


Rules $l = n-1$ and $m$ can be $l$, 0 and $-l$

Orbitals
- $s = l = 0$ (sharp)
- $p = l = 1$ (principal)
- $d = l = 2$ (diffuse)
- $f = l = 3$ (fundamental)
- $g = l = 4$

(terms for each of the orbitals derive from spectroscopy)
BONDING

For the purpose of the discussion on bonding here we will consider the Bohr model as a model of an atom.

The Basic ‘Bohr’ Model of an Atom.

The Bohr model depicts the atom as a small positively charged nucleus surrounded by electrons that travel in orbits around the nucleus, (n=1, n=2, etc.) attraction to the nucleus keep the electrons in these orbits. This is now the accepted model of the atom and was an improvement on the Plum-pudding model of 1904 and the Rutherford model of 1911. In the Bohr model only the outer shell (valance shell) electrons are available to bond with other atoms. A complete outer shell of 8 electrons (or 2 in the case of the innermost shell) is the most energetically favorable.

BONDING

**Covalent:** These bonds occur when atoms ‘share’ electrons in order to follow the ‘octet’ rule, i.e. when they have a full outer shell of 8 electrons (2 electrons in the innermost shell).

Simple covalent bond formation. The atoms ‘share’ electrons to provide a full outer orbital for all atoms. (note for each electron pair - one electron from each atom)

**Ionic:** These bonds occur by electrostatic attraction between two oppositely charged ions one of which is usually a metal ion. In forming an ionic bond the metal donates an electron, due to a low electronegativity to form a positive ion (cation). The counterion gains one or more electrons (anion).
Simple Ionic Bond Formation between Sodium and Chlorine. The chlorine has gained an extra electron from the sodium meaning that both atoms have full and hence stable outer shells.

**Metallic:** In metallic bonding all outer (conduction) electrons of the atoms are in an electron cloud of delocalised electrons. The ‘sea’ of electrons is amongst a lattice of positively charged ions (cations), but are free to move hence allowing conduction.

**Van Der Waals:** This force is due to charge separation forming dipoles in polar molecules. Oppositely charged dipoles on different molecules attract each other forming Van der Waals bonding. This force stops water from being a gas at room temperature with each water molecule being bonded via Van der Waals forces to 4 others!

**SHORT AND LONG RANGE STRUCTURE OF SEMICONDUCTORS**

In Carbon or Silicon (or any other group IV element) the outer shell has 4 electrons occupying the 8 available electron states, these 4 outer electrons are known as ‘valance electrons’. The four remaining states in the outer shell could be thought of as ‘holes’ – energy states that are missing an electron.

It is clear that the physical nature of the atoms that make up a semiconductor is the most important factor in determining the electrical properties. But the other fundamental features which contribute to the electrical properties of a semiconductor need consideration and these are:

1) crystallinity
2) structure
3) defects in composition

**CRYSTALLINITY**

Solids form when the atoms or molecules that make up a compound can no longer escape the intermolecular forces on them and are instead fixed into place. Solid materials are classified by the way the atoms are arranged within the solid.

**Amorphous materials:** Atoms are arranged in a random order. Often produced by rapidly cooling molten material and can also be produced by additives which interfere with crystallization.

Figure 1: Amorphous Material.
**Crystalline Materials:** Atoms are arranged in a highly ordered structure. In general, single crystal material features superior electronic and photonic properties when compared to polycrystalline material and amorphous materials, but it is far more expensive to produce. All high-end semiconductor electronic and photonic materials are fabricated using single-crystal substrates.

![Crystalline Material](image)

Figure 2: Crystalline Material.

**Polycrystalline materials:** Atoms arranged with a high degree of short-range order and no long-range order. This occurs when multiple crystalline grains form independently, grain boundaries are produced between the grains.

![Poly-crystalline Material](image)

Figure 3: Poly-crystalline Material.

**STRUCTURE**

When atoms form regular crystalline materials, they do so in fairly predicable ways, with very predicable arrangement of atoms. The *structure* of a crystal is largely dependent on the number of types of atom, the relative sizes of atoms, the temperature of formations & sometimes the nature of surrounding materials.

There are 12 common types of crystal structure (you could look these up), including various forms of square, rectangular and hexagonal. Silicon adapts the *cubic* “diamond” structure (I will include some details below). GaAs and ZnSe are rectangular (Sphalerite) GaN and ZnO are hexagonal (Wurtzite).
The type of structure of a semiconductor determines the distance between atoms and greatly affects the electronic and optical properties of the semiconductor. Carbon is a great example of a material that comes in a number of forms.

![Types of Carbon Structures](http://graphene.nus.edu.sg/sites/default/files/static_files/allotropes.001_1.jpg)

Diamond lattice (same as Si)

Each of the structures above feature in modern electronics devices, and although each is made out of carbon on its own, each has very different properties.
CRYSTAL STRUCTURE OF SILICON

Crystal planes

The orientations of surfaces of crystals are defined by crystal planes. Look carefully at the figures below showing (100), (110) and (111) crystal planes.

Lattice Constants

The lattice constant of cubic crystal structures (a) is the physical dimension of a unit cell.

Simple Cubic

Face Centred Cubic

Body Centred Cubic

Silicon has diamond structure, which can be considered as a face centred cubic structure with an extra atom at \(a/4 + b/4 + c/4\) – or two inter-penetrating face centred cubic lattice structures.

From www.Fujitsu.com

Lattice Constant (a) is the distance of a cubic unit cell. Note that silicon is composed of tetrahedral silicon and the (100) plane is a pseudo-face centred cubic. (8 corner lattice atoms, 6 face centred atoms, 4 atoms in basis (look above for the figure of the silicon/diamond lattice))

**What is the volume density of silicon which has a lattice constant of 5.43Å?**

\[
\text{(100) plane } = \frac{4 \times 1 + 1}{(5.43 \times 10^{-8})^2} = 6.8 \times 10^{14} \text{cm}^{-2}
\]

Number of atoms per cube (8 corner lattice points, 6 face centred points, 4 atoms in basis)

\[
= \left(8 \times \frac{1}{8} \right) + \left(\frac{1}{2} \times 6 \right) + 4 = 8
\]

Volume density (silicon atoms) = \(\frac{\text{Number of atoms per cube}}{(\text{lattice constant})^3}\) = \(\frac{8}{(5.43 \times 10^{-8})^3}\) = \(5 \times 10^{22} \text{cm}^{-3}\)
**ENERGY LEVELS**

In this section we need examine how atoms within particular crystal structures create *electronic band structures* that ultimately determine the electronic and optical behaviour of semiconductors. We will see that semiconductors, like atoms, have very specific electron energy levels, however whereas the energy levels in atoms are very predictable because atoms are very simple things, the energy levels in semiconductors are predictable because they form crystals with very regular structure.

**PAULI EXCLUSION PRINCIPLE**

Electrons surrounding a single atom have very well defined energies levels (orbits), however. Electrons can only exist at these very discrete energies levels.

The Pauli Exclusion Principle dictates that no two energy levels within a system can be the same, so when two atoms are added together the number of energy levels doubles. Systems with four atoms will again double the number of energy levels. For a system with many atoms it is no longer useful to think of discrete energy levels and it is more useful to think of the system containing continua of energy levels - *energy bands*.

So, within a crystalline material an electronic band-structure is formed consisting of allowed energy bands and forbidden energy gaps, electrons cannot exist within the *forbidden energy gaps*.

Two very important energy bands emerge. The highest energy band filled with electrons (at 0K) is called the **Valence Band**, and the lowest unfilled energy band (at 0K) is the **Conduction Band**.

![Energy Band Formation Diagram](image)
ELECTRONS AND HOLES

We should all be happy with the concept of electrons, however a new concept to most of you will be the ‘electron hole’ or as it is more commonly referred to the ‘hole’.

A hole is simply the absence of an electron from an otherwise full valence band.

It is rather convenient not to think about a hole as an absence of an electron, it is much easier to understand and visualize holes as holes. ‘Particles’ that move around a semiconductor (like electrons) but existing in the valence band instead of the conduction band and of unit positive charge rather than unit negative charge.

SEMICONDUCTOR AT 0K

At 0 Kelvin there is no thermal energy in a system and all the electrons in a system will exist in the valence band of energies. There are no electrons in the conduction band and no holes in the valence band. There is no possibility of electrons moving in the conduction band (because there aren’t any) and no chance of holes moving (because there aren’t any); if nothing can move then the material cannot conduct electricity.

![Semiconductor band diagram at 0K with all electrons in valance band.](image)

SEMICONDUCTOR AT ROOM TEMP

Now if we add temperature, some electrons will have enough thermal energy to move across the energy gap, from the valence band to the conduction band. Now we have electrons in the conduction band and as a direct consequence, the same number of holes in the valence band. More than this, because there are many empty electron states in the conduction band electrons there are free to move about the lattice – they are free electrons. Similarly, because the valence band is mainly full of occupied state the holes are free to move about the lattice they are free holes. These free carriers allow the material to conduct electricity.

(Note: holes move as neighbouring electrons fill them in - leaving behind a hole)

![Semiconductor band diagram at room temp (RT).](image)
SEMICONDUCTORS/ INSULATORS/ METALS

We can now understand the difference between semiconductors, metals and insulators.

**Semiconductors:** The energy gap $E_g$ is $>0$ but is small enough for electrons to feasibly move up from the valance band to the conduction band via external excitation ($E_g < 5$eV).

![Energy Band Diagram of a Semiconductor.]

**Insulators:** The energy required to jump from Valence to Conduction bands is unfeasibly large ($>5$-10eV).

![Energy Band Diagram of an Insulator.]

**Metals and other conductors:** The energy gap $E_g$ is 0 or less (that is the Valence and Conduction bands overlap).

![Energy Band Diagram of a Conductor.]

**BANDGAP ENERGY**

The bandgap energy ($E_g$) is usually given in electron volts ($eV$) as shown below:

Charge $\times$ Voltage = Energy.

and the charge on a single electron ($e$) is $1.6 \times 10^{-19}$ C.

so, $1eV = 1.6 \times 10^{-19}$ C $\times$ 1 V $= 1.6 \times 10^{-19}$ J

The electron volt or $eV$ is a unit of ENERGY and we always use $eV$ for energy.
EXCITATION MECHANISMS

Electrons can be move between bands via either thermal or optical processes.

Thermal Excitation

As previously seen the addition of heat to a semiconductor can provide sufficient energy in a system as to allow some of the valance electrons to be excited into the conduction band.

The thermal energy of an electron is given by \( kT \). Where \( k \) is Boltzmann’s constant \( (1.38 \times 10^{-23} \text{ J} \cdot \text{K}^{-1}) \) and \( T \) is Temperature in Kelvin.

\[
(\text{At room temperature } kT = 26 \text{ meV})
\]

Now the excitation rate of electrons into the conduction band is typically of the form:

\[
\text{Excitation Rate} = \text{const} \times \exp\left(-\frac{E_g}{kT}\right)
\]

That is to say that the number of electron in the conduction band will increase exponentially with temperature.

Optical Excitation

Optical excitation can also excite electrons from the valance to the conduction band.

In this process all the energy of the photon is transferred to the electron, this process can only work if the photon energy is greater than the bandgap. The photon is absorbed in this process and ceases to exist – it is annihilated.

The energy of a photon is given by \( E = h\nu \)

Where \( h \) is Plank’s Constant and \( \nu \) is the frequency of the photon.

It can readily be shown that:

\[
\text{Energy of a photon (in eV)} = \frac{1.24}{\text{wavelength (in } \mu\text{m})}
\]
DIRECT AND INDIRECT SEMICONDUCTORS

There are direct and indirect semiconductors, and it is important that you know the implications of this, though it is not fully necessary to understand the detailed nature of energy momentum diagrams for this module.

We will start by considering an electron in vacuum, followed by an electron in a crystal and the concept of effective mass, then we will look into the detail of energy-momentum diagrams.

Energy and momentum

For electrons in a given system –

1) all energy must be conserved
2) all momentum must be conserved

Energy and momentum are very closely related:

\[ E = \frac{1}{2}mv^2 \quad \text{and} \quad \rho = mv \]

So,

\[ E = \frac{\rho^2}{2m} \quad \text{Eqn. 1} \]

If we were to plot energy against momentum of an electron in vacuum we would get a parabolic curve. If we know the energy so we can work out its momentum and the two values are immutably linked.

Electrons in a crystal and the concept of effective mass of an electron

If we compare an electron in a vacuum as compared to in a crystal, and apply a voltage – for both cases the electron will accelerate. However, the electron in the vacuum will accelerate at a greater rate a compared to in the crystal. The electron in the crystal has to travel through the lattice. Surprisingly free-electron equations can be applied for crystals by simply substituting the mass of an electron in vacuum for the effective mass in a crystal.

An electron in a vacuum will have an energy-momentum plot similar to a semiconductor behaves much the same, but it behaves slightly differently because the crystal structure that surrounds the electron has the effect of modifying the relationship between momentum and energy, by effectively changing the mass of the electron. This whole situation is modelled by considering the electron to have a given effective mass in a given semiconductor.

The same concept can be applied to holes, only holes have a negative effective mass (when plotted on the same diagram as an electron).

So when we plot an energy-momentum diagram for electrons and holes in a semiconductor we get something like this…
Here, momentum is represented by $k$ (actually the *propagation constant* – see later). Here we have plotted a direct bandgap as the minimum of the conduction band (where all the electrons happen to be) is aligned (in terms of momentum) with the top of the valence band (where all the holes happen to be).

**Direct Band-gap Transition**

In a direct band-gap transition the movement of an electron from the valence to the conduction band, via optical excitation, *can take place at the bandgap minimum with no need for momentum change*. Meanwhile, movement of an electron from the conduction to the valence band known as *electron-hole recombination* (another type of *annihilation*) requires the electron to lose energy by emitting a photon of energy $E_g$. See the figure above is a simplified energy-propagation plot for a direct band-gap transition.

**Indirect Band-gap Transition**

An indirect semiconductor bandgap looks like this:

Energy-Momentum diagram of indirect bandgap transition.
In indirect band-gap transitions the electron must change both its momentum and its energy when moving between the conduction and valence bands.

But in order for the electron to change its momentum another particle needs to be involved in order to maintain energy and momentum conservation. The particle which is involved in this indirect transition is called a phonon which in simple terms is a lattice vibration. Lattice atoms at any temperature above 0 K will vibrate around their position within the lattice. These take the form of vibration waves moving throughout the lattice structure.

In an indirect semiconductor the momentum change required for an electron to move across an indirect energy gap is achieved via electron-phonon interaction.

This difference between direct and indirect band-gap materials dictates the efficiency of any optical process within a semiconductor, indirect semiconductors (such as silicon) do not make efficient light-emitting devices they also need to be very thick to make solar cells or photodiodes.

Meanwhile, direct bandgap materials such as GaAs and GaN can be very efficient at producing photons.

Quantum version of energy – momentum diagrams (what $k$ is)……

Now, as we know, electrons are waves with probability distributions

If we think of electrons waves, the relationship between momentum ($\rho$) and de Broglie wavelength ($\lambda$) of the electron is:

$$\rho = \frac{h}{\lambda}$$

Where $h$ is the Planck Constant ($6.626068 \times 10^{-34} \text{ m}^2 \text{ kg} / \text{s}$ (or J/s))

And the propagation constant ($k$) which is parameter similar to momentum

$$k = \frac{2\pi}{\lambda}$$

So by substituting these into Eqn. 1 above, energy of an electron is given by

$$E = \frac{h^2 k^2}{2m}$$

Where, $h = \frac{\hbar}{2\pi}$.

Or

$$E = \frac{h^2 k^2}{2m^*}$$

where $m^*$ is the effective mass of the electron in a crystal.
Thus

\[ \frac{d^2 E}{dk^2} = \frac{\hbar^2}{m^*} \]

Thus the curvature of the $E-k$ parabola defines the effective mass ($m^*$).

The effective mass concept is a way of taking into account the complex interactions of an electron when in a crystal lattice. If we simply change the mass of the electron from its actual vacuum (free electron) mass we can modify the gradient of the $E-k$ parabola.

Energy-Momentum diagram of free electron and electron in a crystal lattice.
Now

The effective mass of the electron and the hole depend upon the direction of travel within the crystal lattice. In general the energy band structure of a semiconductor is much more complex than the diagrams we have discussed.

Below are the energy band diagrams for germanium, silicon and gallium arsenide.

Note silicon, Germanium and Gallium Arsenide all have the same diamond crystal structure, but germanium and silicon are indirect semiconductors and gallium arsenide is a direct semiconductor.

The 3D periodic potential within the crystal is dependent upon direction. Thus, the electron wavelength and crystal momentum $k=2\pi/\lambda$, differs with direction within the crystal lattice. Since true plots of E-k are 4D plots, the band structure is represented as ‘slices’ along ‘directions’ (wave vectors). You will notice that the plots shown above are annotated with the crystal planes (learnt previously in the crystallinity section).

Γ-point is the point where $k=0$
X-point is in the <100> direction
L-point is in the <111> direction

(Note an extension of these concepts (Reciprocal Space Description and Brillouin Zones) is outside the scope of this module.)
CARRIER CONCENTRATIONS

Now we know something about the nature of energy bands in semiconductors and we know a little about how electrons and holes behave in those energy bands. The next important thing to look at is the concentrations of free electrons and free holes in semiconductors, as this will ultimately determine the conductivity of semiconductors.

INTRINSIC SEMICONDUCTORS

An Intrinsic semiconductor is a completely pure semiconductor with no other atoms or impurities present.

In this case for every electron in the conduction band there is a hole in the valence band.

\[
\text{Free electrons} = \text{Free holes}
\]

\[n = p = n_i\]

Where:

- \(n\) is the free electron concentration \((\text{cm}^{-3})\),
- \(p\) is the free hole concentration \((\text{cm}^{-3})\)

and

- \(n_i\) is the intrinsic carrier concentration \((\text{cm}^{-3})\)

(Note: the intrinsic carrier concentration, is a function of temperature that takes different values for different semiconductors and in fact, different crystals of the same semiconductor! – the purity and the perfection of the semiconductor have a big part to play in this and impurities in a crystal at a level of 1 impurity atom for every \(10^6\) silicon atoms can significantly affect the value of \(n_i\) we will use a typical value of \(1.45\times10^{10}\, \text{cm}^{-3}\) in our calculations.)

LAW OF MASS ACTION

The law of mass action is an important concept put simply:

\[n \times p = n_i^2\]

and while it is clear that this will be the case for intrinsic semiconductors it also applies to extrinsic semiconductors.

EXTRINSIC SEMICONDUCTORS

- doped material

The doping process can greatly alter the electrical characteristics of the semiconductor. This doped semiconductor is called an extrinsic material

n-Type Semiconductors (negatively charged electron by adding donor)
p-Type Semiconductors (positively charged by adding acceptor)
DONORS AND ACCEPTORS

In order to control the electron and hole concentrations the semiconductor is usually doped with either electron donor impurities (donors) or electron acceptor impurities (acceptors).

Doped semiconductors are known as extrinsic semiconductors.

Electron Donors

An intrinsic semiconductor can be doped with electron donors to increase the number of free electrons.

This leads to n-type conduction mechanisms and n-type material in which there are many more free electrons than free holes.

![Diagram of donor doping with extra electron]

In the figure a phosphorous atom has been introduced to the Silicon lattice, this is a donor impurity. As long as the phosphorous concentration is small (P/Si = 1/1000), the phosphorous atoms sit on silicon sites forming exactly the same covalent bonds as the silicon atoms they replaced. Phosphorous however, has one too many electrons, these electrons are not required as part of the valence structure and with very little energy separate from the phosphorous atoms to become free electrons – leaving behind positively charged ionised donors.

Electron Acceptors

An intrinsic semiconductor can be doped with electron acceptors to increase the number of free holes.

This leads to p-type hole conduction and p-type materials in which there are many more free holes than electrons.

![Diagram of acceptor doping with missing electron]

In intrinsic silicon, each atom has a full outer shell of electrons.
In the figure a boron atom has been introduced; this is an accceptor impurity. The boron atoms sit on silicon sites forming exactly the same covalent bonds as the silicon atoms they replaced. Boron however, has one too few electrons, and yet the boron atom has to form part of the valence structure and creates a hole. With very little energy the hole can separate from the boron atom to become a free hole – leaving behind negatively charged ionised acceptors.

**Doping of silicon wafers.**

Distribution Coefficient \((k_d)\)

\[
k_d = \frac{C_S}{C_L}
\]

\(C_S\)=concentration of dopant in solid \((\text{atoms/cm}^3)\)

\(C_L\)=concentration in molten liquid \((\text{atoms/cm}^3)\)

**Worked example**

*An ingot of \(10^{16}\) phosphorus atoms/cm\(^3\) is required where a silicon load of 5kg is used for making the ingot. What is the mass of phosphorus that should be added?*

Distribution Coefficient (Phosphorus in Silicon) \(k_d = 0.35\)

Molecular weight phosphorus = 31

Silicon density = 2.33g/cm\(^3\)

Initial phosphorous concentration in melt \(= C_L = \frac{C_S}{k_d} = \frac{10^{16}}{0.35} = 2.86 \times 10^{16}\text{cm}^{-3}\)

Assume because the concentration is so small that the volume of the melt is calculated from the weight of silicon

\[
\text{Volume} = \frac{\text{mass}}{\text{density}} = \frac{5000g}{2.33g/cm^3} = 2146\text{cm}^3
\]

\[
\therefore\ 2.86 \times 10^{16}\text{cm}^{-3} \times 2146\text{ cm}^3 = 6.14 \times 10^{19}\text{ Phosphorous atoms}
\]

\[
\text{Mass} = \frac{\text{Number atoms} \times \text{Molecular Weight}}{\text{Avogadro's Number}} = \frac{6.14 \times 10^{19}\text{atoms} \times 31g/mol}{6.02 \times 10^{23} \text{atom/mol}}
\]

\[= 3.16 \times 10^{-3} \text{ grams of phosphorus to be added to 5kg of silicon}\]
CHARGE NEUTRALITY

If we want to establish concentrations of majority and minority carriers in doped materials then we need to have additional notation for the abbreviations for the carriers and this usually means that we have a subscript term to the abbreviation which defines whether the material is n or p-type.

- In an undoped semiconductor at room temperature there are thermally generated electrons and holes (n and p). In this case \( n_i^2 = n \cdot p \).
- In a semiconductor doped with only donors (\( N_D \)) then the majority carrier in this n-type semiconductor is the electron (\( n_n \)) – and we can usually assume that the concentration of electrons \( n \sim N_D \) and (There will be minority carriers \( p_n \) but these will be present at very low concentrations \( p_n = n_i^2/N_D \) and are important when we consider devices – remember the pn junction and derivation of the diode equation)
- In a semiconductor doped with only acceptors (\( N_A \)) then the majority carrier in this p-type semiconductor is the hole (\( p_p \)) – and we can usually assume that the concentration of electrons \( p_p \sim N_A \) (There will be minority carriers \( n_p \) \( p_p = n_i^2/N_A \))

In a COMPENSATED semiconductor there are two dopants \( N_A \) and \( N_D \), but whether the semiconductor is n-type or p-type depends upon which dopant is in the highest concentration.

- So let’s first assume we have an n-type compensated semiconductor where \( N_D > N_A \), the majority carrier is the electron (\( n_n \)) and the minority carrier is the hole (\( p_n \)) – here the subscript defines the overall type of semiconductor. (Because we have two dopants we will have electrons and holes present as a result of the lower concentration dopant - and thermally generated ones from the semiconductor - lots of carriers.)
- So for a p-type compensated semiconductor where \( N_D < N_A \), the majority carrier is the hole (\( p_p \)) and the minority carrier is the hole (\( n_p \)).
- For all doped samples we are likely to have holes and electrons arising from both the dopant and thermally excited carriers from the semiconductor itself – so do not distinguish where they are from but these are termed \( n^- \) and \( p^+ \)

Now the issue here is – how do we calculate the Fermi Levels (or other parameters in devices) for these compensated semiconductors. We know we have to establish the concentration of the majority carriers and to do these we need to derive the carrier concentration equations

For any system in equilibrium the overall charge must be zero, so the negative charges (electrons and ionised acceptors) must equal the positive charges (holes and ionised donors).

This gives rise to the charge neutrality equation

\[
N_A^- + n^- = N_D^+ + p^+
\]

Where
- \( N_A^- \) is the ionised acceptor concentration.
- \( N_D^+ \) is the ionised donor concentration.
- \( p^+ \), (often simplified to \( p \)) is the thermal-equilibrium concentration of holes
- \( n^- \), (often simplified to \( n \)) is the thermal-equilibrium concentration of electrons

(Remember that an undoped semiconductor at room temperature will have electrons and holes in thermal equilibrium)
CALCULATING MAJORITY CARRIER CONCENTRATION

We are now in a position to calculate the free electron concentrations and free hole concentrations for semiconductors with known doping concentrations.

To calculate the **majority carrier concentration** we can take the Law of Mass Action and the Charge Neutrality Condition and solve simultaneously for n and p.

**Law of Mass Action**

\[ np = n_i^2 \]

**Charge Neutrality condition**

\[ N_A + n = N_D^+ + p \]

Where both donor and acceptors are present

- \( n_n \) and \( p_p \) - the **majority carrier concentrations** (Note sometimes termed \( n_d \) and \( p_a \)) – see descriptions above (you can only have one or the other – depending on the type of SC).
- \( n_p \) and \( p_n \) - the **minority carrier concentrations** (Note sometimes termed \( n_a \) and \( p_d \)).

By definition

\[ N_D^+ = N_D - n_p \]

\[ N_A^- = N_A - p_n \]

Using the charge neutrality condition (see above) we substitute these terms into the charge neutrality condition. We also identify the electrons and holes

\( (N_A - p_n) + n_n = (N_D - n_p) + p_p \) ....eqn 1

Assume

\[ p_n = p_p = 0 \]

Then eqn. 1 becomes

\[ N_A + n_n = N_D + p_p \]

Using the law of Mass Action, this can be solved

So for determining the concentration of electrons \( n_n \) in an \( n \)-type compensated semiconductor

\[ N_A + n_n = N_D + p_p \]

\[ (p_p - N_A) + (N_D - n_n) = 0 \]

Using the law of Mass action \( np = n_i^2 \)

\[ \left( \frac{n_i^2}{n_n} \right) - N_A + (N_D - n_n) = 0 \]

\[ n_n^2 - n_n(N_D - N_A) - n_i^2 = 0 \]

\[ n_n = \left( \frac{N_D - N_A}{2} \right) + \left[ \left( \frac{N_D - N_A}{2} \right)^2 + 4n_i^2 \right]^{1/2} \]

So if we simplify this a little

\[ n_n = \frac{1}{2} \left[ (N_D - N_A) + \left( (N_D - N_A)^2 + 4n_i^2 \right)^{1/2} \right] \]

Note \( N_D > N_A \) (\( n \)-Type)
We can take the same route to obtain the expression for the majority carrier in a p-type compensated semiconductor.

\[ p_p = \frac{1}{2} \left[ (N_A - N_D) + \left[ (N_A - N_D)^2 + 4n_i^2 \right]^{1/2} \right] \]

(note: the negative solution to the quadratics make no physical sense)

However we can make simplifications sometimes in calculating the majority carrier concentration

If you examine the two equations for majority carrier concentrations, you can readily see that if

\[ N_D \gg N_A \gg n_i \quad \text{then} \quad n_n \approx N_D \]

and if

\[ N_D > N_A \gg n_i \quad \text{then} \quad n_n \approx N_D - N_A, \quad p_p = n_i^2 / (N_D - N_A) \]

Or if

\[ R_A \gg N_D \gg n_i \quad \text{then} \quad p_p \approx N_A \]

and if

\[ N_A > N_D >> n_i \quad \text{then} \quad p_p \approx N_A - N_D, \quad n_n = n_i^2 / (N_A - N_D) \]

These approximations are important and will be used very often within this course, whenever we want to know the majority carrier concentration.

*****You should not substitute these simplifications if you are asked to derive the majority carrier concentration equations.

MINORITY CARRIER CONCENTRATIONS

If we want to know the minority carrier concentrations we need to substitute the majority carrier concentration back into the law of mass action.

\[ n_p = \frac{n_i^2}{p_p} \] and

\[ p_n = \frac{n_i^2}{n_n} \]

(note: here we begin to see the importance of the law of mass action, if we add donors we add electrons and if we add electrons our hole concentration must go down. Similarly if we add acceptors the number of electrons must go down. Electrons and holes will cancel each other out.)
FERMI LEVELS and DENSITY OF STATES

We have already seen how it is easy to calculate the (or even approximate) the number of free electrons and the number of free holes in a semiconductor once we know the impurity concentrations. Unfortunately, however, this does not always give us enough information, we need to be able to understand how majority (and more importantly minority) carrier concentrations change as we add voltages to system and this requires us to consider the statistical distribution of electron energies within a system.

Some general probability concepts for understanding the Fermi-Dirac equation

The distribution or probability density functions describe the probability one expects to find an electron occupying the available energy levels in a system. The actual derivation of these functions is beyond the scope of this module. However, we need to understand the notions and that

- the distribution of electrons over the available energy levels has the same probability as any possible distribution.
- The total number of electrons and total energy of the system must be considered

If we consider a hypothetical system shown in the figure below:

Here the system has equi-distant energy levels and each can contain two electrons of opposite spin. The electrons are indistinguishable from each other. At 0K the electrons (for all the atoms in the material) all occupy the lowest possible energy levels (fully occupied to 10eV) but as the temperature is increased electrons move to higher energy levels. For the example shown (E=116eV) above are a number of possible orbital occupancies with the same total energy at a temperature above 0K. At this temperature the probability of an electron occupying higher energy levels than 10eV increases, with the probability of occupancy of the energy level at 11eV over that at 16eV being greater.

I’ve plotted the small amount of data possible from the simplistic energy level diagram above – even with such a small amount of data it is possible to see the probability of each energy level being occupied at a total energy of 116eV (hypothetical). (It is important to note that there is 50% probability of occupancy of an electron is 10.5eV).
THE CONCEPT OF THE ‘FERMI LEVEL’

As we have learnt previously, for semiconductors there are distinct valence and conduction bands separated in energy by the band gap. (which is different to the simplified example shown above).

The *Fermi-Dirac probability density function* provides the probability that an energy level is occupied by an electron (with half-integer spin and obeying the Pauli exclusion principle) that is in thermal equilibrium with a large reservoir:

\[ P(E) = \frac{1}{1 + \exp\left(\frac{E - E_f}{kT}\right)} \]

*Fermi level* – a reference energy level (hypothetical) at which the probability of occupation by an electron is 0.5. \( P(E) = 0.5 \)

As the Fermi level is a reference level which in schematics is drawn between the valence band and conduction band - upon the doping of the semi-conductor.

For semiconductors doped with

- **ACCEPTORS** - then the Fermi level is closer in energy to that of the VALENCE band.
- **DONORS** - then the Fermi level is closer in energy to that of the CONDUCTION band.

**NOTE:** The temperature does not impact upon the energy of the Fermi Level.

At zero degrees Kelvin, there is no thermal energy and all electrons are in the valence band. There is no chance of any electrons being in the conduction band and the Fermi level sits exactly between the two bands.

(note: Although there are no electrons at the Fermi energy, at 0K (for instance) the probability of an electron below the Fermi energy is 1 and the probability of find one above is 0 so the average is 0.5)
As the temperature is increased the probability of finding an electron in the conduction band increases exponentially, this is thermal excitation. The actual position of the Fermi level doesn’t move with temperature. This only occurs (at equilibrium) when acceptor or donor impurities are added.

We can now draw an energy band diagram with 3 energy levels. $E_C$ is the bottom of the conduction band, $E_V$ is the top of the valence band and $E_F$ the Fermi level.

\[
\begin{array}{c}
E_C \\
\vdots \\
E_f \\
\vdots \\
E_V \\
\end{array}
\]

\[
\begin{array}{c}
\text{add } N_A \\
\vdots \\
E_f \\
\vdots \\
\text{add } N_D
\end{array}
\]

In an intrinsic semiconductor the Fermi level will lie exactly between the conduction and valence bands. In n-type semiconductors it will lie closer to the conduction band, in p-type semiconductors it will lie closer to the valence band.

**DENSITY OF STATES**

So we now know something about the distribution of electrons with respect to energies, all we need now is an idea about the number of available electron states. We already have an idea that there are no electron states in the forbidden gap. We know that there are states available in the conduction and valence band. We don’t know much about the distribution of available states in the conduction and valence bands & how that and the Fermi-level can tell us something about carrier concentrations.

As illustrated above, the Density of States in the conduction and valence bands (left) increase as we move to higher or lower energies respectively. If we multiply this function by the Fermi function (middle) we can plot the numbers of electrons and holes at each energy (right). To obtain the carrier concentrations ($n$ & $p$) we can then integrate the number of carriers over all ‘$E$’ in the conduction and valence bands.
In the figure provided, we have placed the Fermi level very close to the valence band (representing a p-type semiconductor), we can see that this has the effect of shifting the Fermi function such that there is a relatively large hole population and a small electron population (though the in the figure this is not to scale and is merely illustrative).

By following these procedures it can be shown that

\[ n = N_c \exp\left(-\frac{E_c - E_f}{kT}\right) \]

and

\[ p = N_v \exp\left(-\frac{E_f - E_v}{kT}\right) \]

Where \( N_c \) is the effective density of states in the conduction band and \( N_v \) is the effective density of states in the valance band.

For a more complete explanation of the calculation of Density of States and Carrier Concentrations please refer to Appendix 1 in the book ‘Introductory semiconductor device physics’ by Greg Parker.

**CARRIER CONCENTRATIONS AND INTRINSIC ENERGY**

Now, instead of worrying about \( N_v \) and \( N_c \) it is much simpler to return to our earlier notion of an intrinsic carrier concentration and to relate that to an intrinsic energy level.

In an intrinsic semiconductor the intrinsic energy is the same as the Fermi energy and found midway between the conduction and valence bands:

\[ E_i = \frac{E_v + E_c}{2} \]

(note: here we are neglecting a small thermal energy term that comprises the true intrinsic energy)

Now we already know that the carrier concentrations at \( E_i \) it is in fact \( n_i \) ! When \( E_f = E_i \) then \( n = p = n_i \).

We can rearrange our Density of state equations above to insert a reference level \( E_i \)

Using this reference value of \( E_i \) it is possible to calculate the carrier concentrations

\[ n = n_i \exp\left(\frac{E_f - E_i}{kT}\right) \]

\[ p = n_i \exp\left(\frac{E_i - E_f}{kT}\right) \]

(Remember we have chosen \( n_i \) for silicon at room temperature to be \( 1.45 \times 10^{10} \text{cm}^{-3} \))

These two equations allow us to:
Find $n_n, p_n, p_p, n_p$ when given $E_f$

or to:

find $E_f$ when given $n_n, p_n, p_p, n_p$ or $(N_D$ and $N_A$)

TEMPERATURE DEPENDENCE OF CARRIERS

Finally for carrier concentrations, we have talked about how the intrinsic carrier concentration changes with temperature, we have also suggested that a certain amount of energy (temperature) is required to ionise impurities. We need to see how these effects combine to provide a picture of carrier concentration variation with temperature. This is best shown diagrammatically:

Free electron concentration for silicon doped with $10^{15}$ donors cm$^{-3}$ as a function of temperature.

This figure shows that silicon has an amazing stability around room temperature.

CARRIER CONCENTRATIONS SUMMARY

$n_n, p_p$ = Majority Carriers (free electron and free hole in n and p–type)

$n_p, p_n$ = Minority Carriers (free electron and free hole in p and n–type)

$N_D$ and $N_A$ donor and acceptor impurity concentrations

$n_i$ = intrinsic carrier concentration

$E_i$ = intrinsic energy

$E_f$ = Fermi energy
For Majority Carriers

\[ n_n = N_D \text{ (or } N_D - N_A) \]
\[ p_p = N_A \text{ (or } N_A - N_D) \]

For Minority Carriers

\[ n_p = \frac{n_i^2}{p_p} \text{ and } p_n = \frac{n_i^2}{n_n} \]

If we know the Fermi level

\[ n = n_i \exp\left(\frac{E_f - E_i}{kT}\right) \]
\[ p = n_i \exp\left(\frac{E_i - E_f}{kT}\right) \]

Band Diagrams

\[ \text{EC} \quad \text{Ef} \]
\[ \text{Ei} \]
\[ \text{Ev} \]

n-type semiconductor \quad p-type semiconductor

EXAMPLE CALCULATIONS

Calculate the position of the Fermi level in the following two n-type semiconductors:

Semiconductor A: \( N_D = 1 \times 10^{17} \text{ cm}^{-3} \)
Semiconductor B: \( N_D = 1 \times 10^{19} \text{ cm}^{-3} \)

Using \( n_n = N_D \) we can say that:

Semiconductor A: \( n_n = 1 \times 10^{17} \text{ cm}^{-3} \)
Semiconductor B: \( n_n = 1 \times 10^{19} \text{ cm}^{-3} \)

Now,

\[ n = n_i \exp\left(\frac{E_f - E_i}{kT}\right) \]

This can be re-arranged to give:
\[ E_f - E_i = kT \ln \frac{n}{n_i} \]

We know the values of \(n\) are \(1 \times 10^{17} \text{cm}^{-3}\) and \(1 \times 10^{19} \text{cm}^{-3}\) for semiconductors A and B respectively.

We also know that at room temp \(kT = 26\text{meV}\) and \(n_i = 1.45 \times 10^{10} \text{cm}^{-3}\).

As such the values of \(E_f - E_i\) can be calculated for the two semiconductors:

For Semiconductor A: \(E_f - E_i = 409\text{meV}\)

For Semiconductor B: \(E_f - E_i = 538\text{meV}\)

This can be shown on a band diagram as below:

**CURRENTS FIELDS AND POTENTIALS**

**ELECTRON AND HOLE MOTION**

Any given electron has a ‘thermal energy’ when at temperatures above 0K. It is this energy which allows electrons to move within a semiconductor. When the semiconductor is in thermodynamic equilibrium (with no electric field present) electrons move randomly within the crystal.

The electron will proceed in any given direction at the thermal velocity until its motion is disturbed by a photon, lattice defect or an impurity within the semiconductor. This interaction will normally be elastic and the sending the electron off in a **random** direction. As such the movement of the electrons due to thermal energy is totally random and therefore there is no net current flow under these conditions.
The thermal energy of a photon is \( kT/2 \) per degree of freedom. In a section of semiconductor the electron will have 3 degrees of freedom \((x,y,z)\) and as such will have a thermal energy of \( 3kT/2 \).

As such electrons in a semiconductor in thermal equilibrium will move at an average speed.

\[
V_{th} = \sqrt{\frac{3kT}{m^*}}
\]

(note: this is derived by equating the thermal energy of the electron \( 3kT/2 \) to the kinetic energy of \( m^* v^2/2 \), where \( m^* \) is the effective mass of the electron)

There are a number of external influences which may cause a flow of electrons?

- Electric Fields
- Diffusion as a result of concentration gradients
- Magnetic Fields (used by the ‘Hall Effect’ or ‘spintronic’ devices).
- Thermal (non-equilibrium via diffusion – but used to measure carrier type)
- Optical absorption (via diffusion)
- Gravity (only really important in cathode ray tubes and possibly space)
- Radio waves (not classically “net flow” but can cause large electron oscillations and can be detected (thankfully))

**FIELDS AND POTENTIALS**

We shall begin with a few key definitions. Electric field \( (\mathcal{E}) \) is the change of potential \( (V) \) with distance.

\[
\mathcal{E} = -\frac{dV}{dx}
\]

Also,

force = charge x field

and

energy = force x distance

Using the above we can gain the equation relating energy and potential:

\[
E = -qV
\]

This is energy gained or lost by transporting a charge \( (q) \) from one place to another where the potential difference is \( V \).

We can now use the above to see how an electric field will affect a semiconductor.
Semiconductor in equilibrium

Band diagrams represent electron energy and this should include the energy state (potential) and energy from any applied field. As such when the field is applied to band diagrams they are tilted to represent potential energy.

On an energy band diagram is useful to imagine, electrons rolling ‘downhill’ towards the applied +ve voltage and the holes floating ‘uphill’ towards the applied -ve voltage.

On an energy plot, the lower energy is the higher +ve voltage \((E = -qV)\).

We call carrier flow due to an electric field carrier drift.

**DRIFT CURRENT**

Now that we have found out that drift is the flow of carriers caused by an applied electric field we can consider how we might calculate drift current.

\[
I_n = J_n A = A E n q \times \text{const}
\]
It is clear that the electron current \( (I_n) \) will be proportional to cross sectional area \( (A) \), free electron concentration \( (n) \) the magnitude of the electric field \( (E) \) and the charge of an electron \( (q) \). In addition there is always bound to be a constant.

Formally,

\[
I_n = J_n A = qA n \mu_n E
\]

and

\[
I_p = J_p A = qA p \mu_p E
\]

where, where \( J_n \) and \( J_p \) are current densities, and \( \mu_n \) and \( \mu_p \) are the electron and hole mobilities, respectively.

Mobility is another material constant for a given semiconductor crystal and a function of temperature, typical values for mobility are provided in the information sheet. The units of mobility are \( cm^2/V\cdot s \).

Mobility is another value that depends upon the quality of the semiconductor crystal, impurities (even doping) and structural defects will reduce mobility values. Often mobility is a figure of merit for a semiconductor and manufacturers are always keen to get the highest values they can.

**SEMICONDUCTOR RESISTANCE**

The drift current equations can be used to derive equations for the resistance and resistivity of a semiconductor.

\[
R = \rho \frac{L}{A}
\]

\[
\rho = \frac{1}{q(\mu_n n + \mu_p p)}
\]

In the above the units of resistivity are \( \Omega \text{cm} \).

A full derivation of the resistivity equation can be found in the core text.

Meanwhile conductivity \( (\sigma) \) is \( 1/\rho \)

**DIFFUSION CURRENT**

Diffusion is the process of particles moving from regions of high concentration to regions of low concentration. This process will eventually lead to a uniform distribution of particles once equilibrium is reached. Diffusion does not require external forces and the particles moving about using only thermal motion. If these particles are charge carriers as they move around they take charge with them. The movement of charge will result in a *diffusion current*.

The difference between drift current and diffusion current is that drift current depends on the electric field applied and as such if there's no electric field, there's no drift current. Diffusion current occurs even though there isn't an electric field applied to the semiconductor.
It can be seen from the diagram above that if we begin with a large potential gradient the concentration of particles will diffuse to reduce this gradient until we have a gradient of 0. It is this concentration gradient change which leads to carrier flow.

\[ I_n = J_n A = qA \frac{dn}{dx} \times \text{const} \]

Where, \( I_n \) is the electron diffusion current, \( J_n \) is the current density and \( \frac{dn}{dx} \) is the electron concentration gradient.

Formally the two separate standard equations for the electron and hole diffusion currents are:

\[ I_n = AqD_n \frac{dn}{dx} \]
\[ I_p = -AqD_p \frac{dn}{dx} \]

In the above \( D_n \) and \( D_p \) are the **Diffusion constants** with unit cm\(^2\)/s.

**MOBILITY AND DIFFUSION (µ,D)**

Mobility and diffusion are related by the Einstein Relation which is simply:

\[ D = \frac{\mu kT}{q} \]

When calculating the electron mobility or diffusion you must use \( D_n \) and \( \mu_n \) and when calculating hole mobility or diffusion you must use \( D_p \) and \( \mu_p \).

Mobility can also be calculated using the following equation:

\[ \mu = \frac{q \tau}{m^*} \]
As seen in the Einstein Relation, mobility is a function of temperature.

![Mobility as a function of temperature.](image)

Mobility as a function of temperature.

**CARRIER TRANSPORT (IN SUMMARY)**

- The net flow of electrons and holes in a semiconductor will generate currents.
- The process by which these charged particles move is called transport.
- The two basic transport mechanisms in a semiconductor crystal are:
  - **DRIFT**: the movement of charge due to electric fields
  - **DIFFUSION**: the flow of charge due to density gradients

**Worked example**

*Calculate the drift current (density) in a semiconductor for a given electric field. Consider a germanium sample at T=300K with a doping concentration of N_d=0 and N_a=10^{16} \text{ cm}^{-3}. Assume complete ionisation and the electron and hole mobilities are 3900 \text{ cm}^2/\text{V.sec} and 1900 \text{ cm}^2/\text{V.sec}. The applied field is E=50V/cm.*

Since N_a>N_d, the semiconductor is p-Type and the majority hole concentration

\[ p_0 = \frac{1}{2} \left[ (N_a - N_d) + \left( (N_a - N_d)^2 + 4N_i^2 \right)^{1/2} \right] \approx 10^{16} \text{ cm}^{-3} \]

thus \( p_0 \approx N_a \)

The minority carrier electron concentration is

\[ n = n_i^2/p = (2.4 \times 10^{13})^2/10^{16} = 5.76 \times 10^{10} \text{ cm}^{-3} \]

For this extrinsic p-type semiconductor

\[ J_{\text{drift}} = J_p + J_n = q(\mu_p n + n\mu_p)\varepsilon \approx qN_a\mu_p\varepsilon \]

\[ J_{\text{drift}} = (1.6 \times 10^{-19})(1900)(10^{16})(50) = 152 \text{ A/cm}^2 \]
**PN JUNCTIONS**

We traditionally start by considering two separate regions of semiconductor, one region is n-type and the other p-type. We can draw the band diagrams for each region side by side and then consider what would happen if the two regions were magically joined together (in practise, this is not possible).

When the two regions are joined the relative position of the Fermi levels indicate a big difference in carrier concentrations on either side of the junction. This is a carrier concentration gradient and will cause diffusion currents until and equilibrium is reached.

In this non-equilibrium situation

- electrons **diffuse** from **n** to **p** and holes **diffuse** from **p** to **n**.
- areas of fixed charge (ionised impurities) are left behind forming **depletion regions**
- these regions of fixed charge produce a field that changes the energy alignment between the n and p region, creating a **built-in potential**
- this change in potential creates a drift currents in which electrons to **drift** from **p** to **n** and holes to **drift** from **n** to **p**.
- eventually the diffusion and drift currents cancel each other out and an equilibrium is reached
- at equilibrium the Fermi level is flat throughout the device and the **built-in voltage** is established in the band-diagram
So, in a pn junction at equilibrium:

- the Fermi level is flat
- $V_{bi}$ is formed as a result of the charge due to ionised impurities in the depletion region
- Drift and Diffusion currents cancel out
- $W_p N_A = W_n N_D$ (charge neutrality)
BUILT-IN VOLTAGE

It is useful to derive an expression for the built-in voltage of the junction, this is readily achieved by considering the positions of the Fermi-levels relative to the intrinsic level in the neutral regions.

We already know that

$$\Delta V_n = \frac{kT}{q} \ln \frac{N_D}{n_i}$$

and

$$\Delta V_p = \frac{kT}{q} \ln \frac{N_A}{n_i}$$

$$\Delta V_n + \Delta V_p = V_{BI}, \text{ so}$$

$$V_{BI} = \frac{kT}{q} \ln \frac{N_D N_A}{n_i^2}$$
PN JUNCTIONS UNDER FORWARD AND REVERSE BIAS

In the following the p-type region has $N_A = 1 \times 10^{17} \text{ cm}^{-3}$ and the n-type region $N_D = 3 \times 10^{17} \text{ cm}^{-3}$ (NOTE the doping level of the n-type region is three times that in the p-type region – thus $W_p$ is three times the width of $W_n$ (see later))

**Equilibrium:**

Electron drift occurs from p-type to n-type and electron diffusion occurs from n-type to p-type. In contrast holes drift from n-type to p-type and hole diffusion occurs from p-type to n-type. There is no net flow of current.

**Forward Bias:**

Under forward bias the potential barrier to carrier movement is reduced along with the width of the depletion region ($W$). This causes an increase in diffusion currents and reduces drift currents.
Reverse Bias

Under reverse bias the potential barrier is increased along with the width of the depletion region. This leads to a decrease in the diffusion currents as the potential barrier reduces the concentration gradients and increasingly favours the movement of electrons from the p-type region and holes from the n-type region.

Meanwhile, the drift currents remain more or less the same and still very small. Drift currents rely on the movement of electrons from the p-type region and holes from the n-type region, and although the field across the junction is increased by a reverse bias there remains a very low concentration of minority carriers. Under reverse bias only small currents, largely independent of reverse bias voltage flow through pn-junctions.

In this way, we can see that a pn junction can allow large currents in forward bias and negligibly small currents in reverse bias, and act as a diode.

**POISSONS EQUATION**

Poisson’s equation is generally expressed as:

\[
\frac{\delta \mathcal{E}}{\delta x} = -\frac{\delta^2 V}{\delta x^2} = q\left(\frac{p^+ + N_D^+ - n^- - N_A^-}{\varepsilon_s}\right)
\]

Where, \(\mathcal{E}\) is the field, \(V\) is potential and \(\varepsilon_s\) is the permittivity of silicon and given by the permittivity of vacuum, \(\varepsilon_0\) multiplied by the relative permittivity of silicon \(\varepsilon_R\).

By applying Poisson’s equation to different regions of the pn-junction we can establish important relationships between the built-in voltages, the depletion region widths, carrier/impurity concentrations, electric field and charge distributions. The first step is to draw a succession of diagrams showing charge, field and potential as a function of distance. This allows us to maintain a clear physical understanding of the junction whilst applying appropriate mathematics.
We assume that the space-charge layer is completely depleted of free carriers, doping is uniform, the junction is abrupt and all of the built-in voltage is dropped across the space charge layer. We can gather that the electric field outside the depletion region is zero and the regions outside the depletion region are charge neutral.

We can see that there are 4 distinct regions within the junction that should be considered. The n and p-type neutral regions and the n and p-type depletion regions.

Using Poisson’s equation it is possible to calculate the field, the potential and the energy within any region as a function of distance. We can extend Poisson’s equation to consider three-dimensional devices; it could even take into account non-equilibrium carrier concentrations for accurate modelling.

For now, all we are interested in is calculating the values for charge, field and potential at the junction and the edges of the depletion regions. A rigorous mathematical approach to finding suitable expressions involves integrating within each of the four regions and applying appropriate boundary
conditions in each region. However, the expressions can also be derived by considering the diagrams for charge, field and potential and applying simple geometry (finding areas) or integrating between limits.

Meanwhile, it is worth remembering that charge neutrality must be maintained and hence:

\[ W_p N_A = W_n N_D \]

At equilibrium the only ‘fixed’ charges are ionised impurities in the depletion regions, so;

In the p-type depletion region:

\[ \frac{\partial \varepsilon}{\partial x} = -\frac{q N_A}{\varepsilon_s} \]

and in the n-type depletion region:

\[ \frac{\partial \varepsilon}{\partial x} = \frac{q N_D}{\varepsilon_s} \]

The maximum field \( \varepsilon_{\text{max}} \) can be calculated by integrating from to \(-W_p\) to 0 or from 0 to \(W_n\) (or multiplying gradient by distance).

\[ \varepsilon_{\text{max}} = q \frac{W_p N_A}{\varepsilon_s} = q \frac{W_n N_D}{\varepsilon_s} \]

\( V_p \) and \( V_n \) in the potential distribution can be found by calculating the area under the field distribution, (or by integrating \(-W_p\) to 0 and from 0 to \(W_n\)) for the two regions:

\[ V_p = \frac{q N_A W_p^2}{2 \varepsilon_s} \quad (1) \]

\[ V_n = \frac{q N_D W_n^2}{2 \varepsilon_s} \quad (2) \]

and

\[ V_{bl} = V_p + V_n \]

From (1) and (2) we obtain the ratio \( V_n/V_p \)

Therefore

\[ \frac{V_n}{V_p} = \frac{N_D W_n^2}{N_A W_p^2} \]

As \( N_A W_p = N_D W_n \) the relationship can also be written as

\[ \frac{V_n}{V_p} = \frac{W_n}{W_p} = \frac{N_A}{N_D} \]

Notes for ELEC1205  Copyright 2016 © University of Southampton
EXPRESSIONS DERIVED FROM POISSON’S EQUATION

Using Poisson’s equation it is possible find expressions for many features of the pn-junction:

Built-In voltage (in terms of impurity concentrations and depletion widths):

\[ V_{BI} = \frac{qN_A W_p^2}{2\varepsilon_s} + \frac{qN_D W_n^2}{2\varepsilon_s} \]

The widths of the two depletion regions can also be derived:

\[ W_p = \frac{1}{N_A} \sqrt{\frac{2\varepsilon_s N_D N_A V_{BI}}{q(N_D + N_A)}} \]

\[ W_n = \frac{1}{N_D} \sqrt{\frac{2\varepsilon_s N_A N_D V_{BI}}{q(N_D + N_A)}} \]

then the total width of the depletion region is:

\[ W_T = W_p + W_n = \frac{\sqrt{2\varepsilon_s (N_A + N_D)}}{q N_A N_D} V_{BI} \]

(you can use this expression to calculate \( W_p \) and \( W_n \) because \( W_p N_A = W_n N_D \))

For a \( p^+ n \) junction this can be approximated as:

\[ W_T = \sqrt{\frac{2\varepsilon_s V_{BI}}{q N_D}} \]

For a \( n^+ p \) junction this can be approximated as:

\[ W_T = \sqrt{\frac{2\varepsilon_s V_{BI}}{q N_A}} \]

We have already seen that:

\[ \varepsilon_{max} = q \frac{W_p N_A}{\varepsilon_s} = q \frac{W_n N_D}{\varepsilon_s} \]

In the above \( V_{BI} \) may be replaced by \((V_{BI} + V_R)\) or \((V_{BI} - V_R)\) to take into account non-equilibrium conditions (applied voltage).
The key to understanding the derivation of the diode equation is an appreciation of the importance of the minority carrier concentrations on either side of the junction

- at equilibrium, diffusion and drift components of current cancel out and there is no net flow of current
- when applying a forward or reverse bias, the balance between the diffusion and drift components of hole and electron current is altered
- when the change in potential across the pn-junction is the built-in voltage, the probability of finding electrons (or holes) on either side of the junction is equal (the Fermi-levels are aligned)
- by adding a forward bias the Fermi levels of the n-type region is raised relative to the Fermi level of the p-type region. This increases the probabilities of finding electrons (& thereby free electron concentration) at the edge of the depletion region of the p-type region and increases the probability of finding holes (& thereby free hole concentration) at the edge of the depletion region in the n-type region
- These large minority carrier concentrations on either side of the junction lead to concentration gradients and large diffusion currents
- Meanwhile the drift components remain small and more or less constant because the minority carrier concentrations are small and remain more or less constant
- as the minority carriers diffuse through the neutral regions they recombine with majority carriers, at a number of diffusion lengths from the junction most of the minority carriers have recombined and the current through the device is carried by majority carriers (that feed the recombination process)

The diode equation is most commonly derived by considering the minority carrier concentrations and current at the edge of the depletion region. So, our first concern, in the derivation that follows is establishing expressions for the minority carrier concentrations as a function of the applied voltage.

**ASSUMPTIONS**

- the junction is abrupt and depletion region edges are abrupt
- the carrier densities are related to the potential across the junction
- injected densities are small in comparison to the majority carrier densities
- there is no generation or recombination
- the majority carrier concentration = the doping concentration

**DEFINITIONS**

\( n_{no} \) and \( n_{po} \) are equilibrium electron concentrations in n & p regions
\( p_{no} \) and \( p_{po} \) are equilibrium hole concentrations in n & p regions
\( p_n \) and \( p_p \) are non-equilibrium hole concentrations at the boundaries of the depletion regions
\( n_n \) and \( n_p \) are non-equilibrium electron concentrations at the boundaries of the depletion regions
STEADY STATE INJECTION (AND CONTINUITY EQUATIONS)

The one dimensional continuity equations for minority carriers can be used to derive expressions that describe the variation of injected minority carrier concentrations with distance. Thus we will start with these equations in deriving the diode equation.

We will consider holes diffusing into to neutral n-type region of a forward biased pn-junction

The minority carrier continuity equation for holes is

\[
\frac{\partial p_n}{\partial t} = -p_n \mu_p \frac{\partial E}{\partial x} - \mu_p E \frac{\partial p_n}{\partial x} + D_p \frac{\partial^2 p_n}{\partial x^2} + G_p - \frac{p_n - p_{n0}}{\tau_p}
\]

drift \hspace{1cm} \text{diffusion} \hspace{1cm} \text{generation} \hspace{1cm} \text{recombination}

\[\tau_p = \text{hole lifetime (s)}\]

Thankfully, this can be simplified

- if we consider steady state conditions, i.e. there is no change in the minority carrier distribution with time.
- there is no field in the neutral region – so no drift
- and there is no carrier generation

thus

\[0 = D_p \frac{\partial^2 p_n}{\partial x^2} - \frac{p_n - p_{n0}}{\tau_p}\]

rearrange and substitute \((p_n-p_{n0})=\delta p\) and the diffusion length \(L_p = \sqrt{D_p \tau_p}\),

therefore, the steady state diffusion equation for holes becomes

\[
\frac{d^2 \delta p}{dx^2} = \frac{\delta p}{L_p^2}
\]

The solution to the diffusion equation for holes has the form

\[
\delta p(x) = C_1 e^{x/L_p} + C_2 e^{-x/L_p}
\]

\(C_1\) and \(C_2\) can be evaluated from the boundary conditions, at large \(x\) the excess hole concentration becomes zero, and at \(x = 0\) it is \(\Delta p\) (the hole concentration at the depletion layer edge), thus

\[
\delta p(x) = \Delta p e^{-x/L_p}
\]

describes the distribution of injected holes.
So for a pn-junction (balanced doping) the minority carrier concentration each side of the depletion region will vary as a function of distance ($x$).

The gradient of this distribution at any point will give rise to a diffusion current,

$$
\delta p(x_n) = \Delta p_n e^{-x_n/L_p}
$$

so if:

$$
\frac{d\delta p(x)}{dx} = -\frac{1}{L_p} \Delta p e^{-x/L_p} = -\frac{1}{L_p} \delta p(x)
$$

and since

$$
\frac{d\delta p(x)}{dx} = \frac{dp}{dx}
$$

The hole diffusion current at distance $x$ from the depletion region is:

$$
J_p(x) = -qD_p \frac{dp(x)}{dx} = q \frac{D_p}{L_p} \delta p(x)
$$

Now, although this current decreases with distance from the edge of the depletion region edge, it is gradually replaced by an increasing electron current flow and the total current due to injected carriers remains the same at any point in the device (i.e. the recombination of holes is pulling in extra electrons).

Thus, the current due to injected holes can be most readily found by considering the concentration gradient of holes at the depletion edge (when all the current is hole current). At $x=0$ (the depletion edge) we have defined

$$
\delta p(0) = \Delta p
$$
So next we need to find an expression for $\Delta p$

We start with our expression for built-in potential:

$$V_{bi} = \frac{kT}{q} \ln \frac{N_i N_D}{n_i^2}$$

This expression was derived from the neutral region majority carrier concentrations, and can in fact be rewritten:

$$V_{bi} = \frac{kT}{q} \ln \frac{p_{p0} n_{n0}}{n_i^2}$$

so

$$V_{bi} = \frac{kT}{q} \ln \frac{p_{p0}}{n_{n0}}$$

re-arranging

$$p_{p0} = p_{n0} e^{qV_{bi}/kT} \quad (1)$$

This relation still holds if we apply a voltage, but the carrier concentrations at the edge of the depletion region are no longer equilibrium values, so

$$p_p = p_n e^{q(V_p-V)/kT} \quad (2)$$

(for forward bias)

Now under low injection conditions $p_p \sim p_{p0}$, and dividing (1) by (2) gives:

$$p_n = p_{n0} e^{qV/kT}$$

And the excess hole density at the edge of the depletion region ($\Delta p$) is given by:

$$\Delta p = p_{n0} \left(e^{qV/kT} - 1\right)$$

Now we know that the total current density due to injected holes is given by

$$J_p(x) = q \frac{D_p}{L_p} \Delta p$$

So,

$$J_p = \frac{qD_p p_{n0}}{L_p} \left(e^{qV/kT} - 1\right)$$

Is the full expression for the current density in the diode due to the injection of holes. A similar procedure finds the expression for the injected electron contribution, which when added gives the diode equation

$$J = J_S \left(e^{qV/kT} - 1\right)$$
where,

\[ J_s = \frac{qD_p P_{n0}}{L_p} + \frac{qD_n n_{p0}}{L_n} \]

Where \( J_s \) is the saturation current density.

\[ I = J A \]

where \( A \) = cross-sectional area

**DIODE CAPACITANCE**

**JUNCTION/DEPLETION CAPACITANCE**

Capacitance is the term for the gradient of charge/voltage and within a pn-junction there are two forms of charge that change with voltage.

We have already seen that the depletion regions of a pn-junction are regions of charge, and we have seen that as we change the voltage across the depletion regions the width and charge change. The **depletion capacitance** or **junction capacitance** results from this effect.

You could (in principle) find an expression for the charge in the depletion region as a function of voltage, and you could differentiate to show that the depletion capacitance \( (C_J) \) is given by

\[ C_J = \frac{\varepsilon_s A}{W} \]

There is a strong analogy with the equation for a parallel plate capacitor, in which the depletion region width \( (W) \) corresponds with the plate separation of a conventional capacitor.

This formula is most important/accurate for reverse bias, and is generally good for low forward bias up to half the built-in voltage

**DIFFUSION CAPACITANCE**

A second instance of charge changing with applied voltage exists under forward bias conditions. Here, the steady-state charge due to excess minority carrier concentrations will increase with
increasing forward bias. This gives rise to diffusion capacitance and begins to be relevant when the forward bias goes beyond half the built-in voltage.

**SOLAR CELLS AND PHOTODIODES**

Current and Voltage in an Illuminated Junction

- A supply of minority carriers from each side of a junction is generated by the photo-excitation of electron-hole pairs
- the resulting current is the generation current and its magnitude depends on the rate of generation of electron-hole pairs
- minority carriers generated in the depletion region or within a an average diffusion length of each side of the depletion region can be swept to the other side of the junction by the built-in electric field

The resulting current due to collection of these optically generated carriers by the junction is

\[
I_{\text{OP}} = qA g_{\text{op}} \left(W + L_p + L_n\right)
\]

Where \(g_{\text{op}}\) is the optical generation rate per unit volume. This current subtracts from the total current through a \(pn\)-junction and the diode equation becomes

\[
I = I_s \left(e^{\frac{qV}{kT}} - 1\right) - qA g_{\text{op}} \left(W + L_p + L_n\right)
\]

Thus, the diode characteristic drops by the optically generated current

The appearance of a forward voltage across an illuminated junction is known as the photovoltaic effect.

Depending on the intended application devices can be operated in either the third or fourth quadrants of its \(I-V\) characteristic

- power is delivered to the device from the external circuit when current and voltage are both positive or both negative (first or third quadrants)
• in the fourth quadrant the junction voltage is positive and the current is negative – power is delivered to the external circuit (in the fourth quadrant the current flows from the negative side of V to the positive side – as in a battery).

If power is to be extracted from the device (i.e. a solar cell) the fourth quadrant is used.

In photo-detecting applications the junction is usually reverse biased and operation is in the third quadrant.

SOLAR CELLS

The basic design features of a silicon wafer solar cell are illustrated below.

• Photons with energy \((h\nu)\) greater than the bandgap \((E_g)\) generate electron-hole pairs
• The Generated minority carriers can diffuse a diffusion length to the edge of the depletion region and then drift across the junction
• The result of this process is an extra electron in the n-type material and an extra hole in the p-type material – the electron can do useful work in an external circuit before recombining with a hole.

Only electron-hole pairs generated within the active thickness of a device can contribute to the current. On average electron-hole pairs generated more than an average diffusion length from the depletion region edge will recombine before reaching the junction region.

A solar cell needs to be designed so that most of the available light is absorbed within the active-thickness. In general high-energy UV and blue light is observed very close to the surface and red and infrared is only absorbed slowly through the whole thickness of the device.
OUTPUT CHARACTERISTICS

We are concerned with the fourth quadrant portion of a diode characteristic, for convenience the reverse current $I_R$ is plotted upwards.

- $V_{oc}$ and $I_{sc}$, the open circuit voltage and the short circuit current are determined by the light intensity and the cell properties.
- Light intensity is often standardised for particular applications, the air mass index is one such standard and is based on the number of atmosphere widths sunlight has to traverse AM0 is for satellites, AM1 applies to the equator, AM2 is 600 latitude, etc.…
- The maximum power that can be delivered to a load by the solar cell occurs when $VI$ is a maximum
- The maximum delivered power ($I_m \times V_m$) is shaded in the figure above
- The ratio of $(I_m \times V_m) / (I_{sc} \times V_{oc})$ is called the fill factor and is a figure of merit for solar cell design
- The efficiency of a solar cell is simply the power delivered as a percentage of the optical power incident on the device
- At a sunny location at the equator about 1kW/m² of irradiation is available (for 12 hours a day)
- By the time you get to the UK at 60° latitude the sunlight, on average has to traverse 1.5 atmospheres further and power available is reduced accordingly. Power is reduced even further if cloudy days are more common, though diffuse light, means you don’t have to have fancy mechanics that allow your cell to track the sun as it travels through the sky
- Efficiencies are typically around 18%, heroic efforts (expensive! but worth making for satellites) can get that up towards 25% for single junction devices

EFFICIENCY LOSSES

1) The photovoltaic effect alone reduce maximum efficiencies for single junction devices to ~30%.

- Photons with $h\nu < E_g$ cannot create electron-hole pairs
- Photon with $h\nu > E_g$ can only generate a single electron-hole pair

2) Reflection – from the front surface

3) Shadowing – of the surface metal contacts
4) Recombination

- Carriers generated outside the active thickness
- Recombination at defects (reduces active thickness)
- Surface recombination

5) Transmission – light that goes straight through

6) Resistance effects (these tend to impact the fill-factor)

- Of semiconductor (impacts drift of majority carriers)
- Of metal contacts and metal-semiconductor contacts
- Shunt resistances (that short circuit the device)

DESIGN CRITERIA

Clearly, design criteria will depend on the particular application in mind. There are however, some general features of solar cell design.

- utilise the maximum amount of available optical energy
- you should collect as much light as possible and minimise losses at every point in the system
- best material quality provides longest carrier diffusion lengths
- the surface of a cell is coated with anti-reflective materials and materials that reduce surface recombination
- the total thickness of the device should be such that all photons are absorbed
- it is desirable to make the junction potential as large as possible, so heavy doping is useful as long as this doesn’t reduce carrier lifetimes
- it is important that the series resistance is small so that power isn’t lost as heat in the device
- The contact to the thin n-type region requires careful design; contact cannot be made along the edge because the thin layer is highly resistive through the plane. So, contact has to be made by distributing an array of metallic fingers that reduce series resistance without interfering appreciably with the incoming light.

GENERAL

- the voltage that can be supplied by a solar cell is generally less than the contact potential, and that is generally less than the band-gap voltage (<1V for Si)
- the current generated depends on the on the illuminated area, and the photon flux (~10 – 100 mA for 1cm²)
- Although 1 V and 100 mA isn’t a great deal of power a large number of such devices can generate significant power
PHOTODETECTORS

The fundamental mechanisms of solar cells and photodetectors are the same. However photodetectors are generally operated with a reverse bias across the pn-junction.

The design parameters for the two devices are also different, notably solar cells need to be matched to the whole solar spectrum, photodetectors are often only tuned to be used for detection of a narrow range of wavelengths.

The basic requirements of a photodetector are:

- Sensitivity
- Efficient conversion of photons to electrons
- Fast response (thus the devices are as thin as possible)
- Low noise
- Sufficient area (usually for coupling an optical fibre)
- High reliability
- Low cost

By highly doping the p-type region, the depletion region penetrates well into the n-type region. The negative bias will further extend the width of the depletion region.

The pn junction photodiode under reverse bias
The generated photocurrent is dependent upon the number of EHP (electron hole pairs) generated in
the depletion region and the drift velocity. There is no injection of carriers, but the bias allows
excess carriers to be delivered providing a current \( I_{\text{ph}} \). For photodetectors speed is often all
important, so it help to make the devices as thin as possible. It also helps to only detect carriers
generated in the depletion region only.

The key application of photodetectors is to convert an optical signal into an electrical signal (i.e.
optical communications – i.e. An area of technology where the University of Southampton
(Optoelectronics Research Centre) have invented many of the key technologies.)

---

Schematic of the components and requirements for optical communication. The information is
converted to a modulated optical signal (i.e. laser source) at the transmitter. This is transmitted
along an optical fibre. At the receiver, and following optical signal processing the optical signal
(temporally modulated (with respect to time)) is detected (photodetector) and the photodetector
output is processed to provide the information transmitted.
BIPOLAR JUNCTION TRANSISTORS

Doping profile and band diagrams

(n⁺pn BJT)

In normal (active mode) of operation

- the emitter-base junction is forward biased
- the base-collector junction is reverse biased

The forward biased emitter-base junction injects minority carriers into the “base” region, once in the base they either recombine with majority carriers or they reach the base-collector depletion region and get swept into the collector.

n⁺pn BJT (Equilibrium)

n⁺pn BJT (Forward Active mode)

n⁺pn BJT band diagrams at equilibrium and forward-active mode.
TRANSISTOR ACTION

The key aspect to the transistor is the width of the base

- carriers injected into the base become the minority carrier in the base region
- in the base the minority carriers have a short lifetime and will recombine readily
- however, if the base is narrow enough then injected carriers will diffuse across the base to the depletion region of the reverse biased base-collector junction, there they will be swept across the junction
- if the base region is too wide then no injected carriers reach the collector, no current flows & you have two diodes back-to-back
- the base width should therefore be shorter than the minority carrier diffusion length in the base

COMPONENTS OF CURRENT

For a well designed transistor (with a heavily doped emitter and a narrow base), the emitter and collector currents are very similar and a small base current is the result of carrier injection from base to emitter.
This understanding of the current flow provides a good explanation of transistor action but represents an over-simplification; there are in fact several other current components that can be important under certain conditions.

An understanding of the pn-junction gives us a clue to the other current components. We would also expect small currents due to the drift of minority carriers across emitter/base and base collector junctions but these are negligible in all practical situations and ‘recombination’ currents – particularly in the base are more significant.

**COMMON EMITTER CURRENT GAIN (β)**

The most important transistor parameter is the common emitter current gain (β). This is the ratio of collector current to base current and is given by:

\[ \beta = \frac{I_C}{I_B} \]

In a well designed transistor the base is sufficiently narrow to ensure that the most injected carriers reach the collector. Thus, in a typical commercial transistor the collector current is approximately one hundred times larger than the base current.

**DERIVATION OF EXPRESSION FOR GAIN**

We can derive equations for the collector and base current for a high speed Si bipolar transistor with a narrow base and a shallow emitter

- The assumption of a narrow base gives an electron distribution in the base that is linear
- The assumption of a shallow emitter gives a hole distribution in the emitter that is linear

With the above assumptions, the equations for the electron and hole diffusion current can be written as:

\[ J_n = qD_{nb} \frac{dn}{dx} \]
\[ J_p = -qD_{pb} \frac{dp}{dx} \]
So to find expressions for the base and emitter currents we need to find the values for the minority carrier concentrations at the edges of the emitter-base depletion region. (just like we did for the diode equation).

If we assume a narrow base and a shallow emitter, we can divide, these values by $W_B$ and $W_E$ to find the gradients and thereby the currents.

Thus, it can be shown that

$$I_B = \frac{qAD_{pe}n_i^2}{W_E N_{DE}} \exp \frac{qV_{BE}}{kT}$$

and

$$I_C = \frac{qAD_{nb}n_i^2}{W_B N_{AB}} \exp \frac{qV_{BE}}{kT}$$

### CURRENT GAIN

The current gain is simply given by the ratio of the collector current to the base current

$$\beta = \frac{D_{nh}W_E N_{DE}}{D_{pe}W_B N_{AB}}$$

The above equation for the current gain allows us to consider the design options for bipolar transistors

- For high gain we need a highly doped emitter and a lowly doped base
- But, if the base doping is too low we get a high base resistance, which will degrade the high frequency operation
- For a high gain we need a narrow basewidth. This will also give high frequency performance
- But a narrow basewidth will also give a high base resistance unless we increase the base doping
- For a high gain we need a deep emitter (but our model assumes a shallow emitter)

The above equation doesn’t apply for deep emitters because our assumption of a linear hole distribution in the emitter breaks down. For a deep emitter the gain is given by:

$$\beta = \frac{D_{nh}L_{pe}N_{DE}}{D_{pe}W_B N_{AB}}$$

(where $L_{pe}$ is the hole diffusion length in the emitter)
**Worked example**

**Question:** Design a pnp transistor with a common-emitter current gain of 50?

Since equation on information sheet is for npn BJT this needs to be revised to

\[
\beta = \frac{D_{nB}N_{DB}W_E}{D_{pE}N_{AB}W_B}
\]

\[
\frac{D_{pB}}{D_{nE}} = \frac{\mu_p}{\mu_n} = 0.35 \quad \text{(Einstein relation)}
\]

All physically reasonable values must be used.
Carrier concentration of emitter should be >> base
For example, \(N_{AE} = 5 \times 10^{18} \text{ cm}^{-3}\) then \(5 \times 10^{17} \text{ cm}^{-3}\).
Let’s say \(W_E = 2 \mu\text{m}\) and \(W_B = 1 \mu\text{m}\) (both considerably smaller than the diffusion lengths).

**So for a Gain of 50, \(N_{DB} = 17 \times 10^{16} \text{ cm}^{-3}\)**

**Question:** A silicon bipolar junction transistor has a 1 \(\mu\text{m}\) emitter with a uniform donor concentration of \(5 \times 10^{18} \text{ cm}^{-3}\), a 100 nm wide base with a uniform acceptor concentration of \(1 \times 10^{18} \text{ cm}^{-3}\) and a collector with a donor concentration of \(5 \times 10^{17} \text{ cm}^{-3}\)

- Therefore this is an npn BJT….

On the information sheet - the gain equation is for a npn BJT

\[
\beta = \frac{D_{nB}W_E N_{DE}}{D_{pE} W_B N_{AB}}
\]

**(NOTE you should show working including using Einstein relation when answering a question)**

\[
\beta = 140.6
\]

Note in this example we are not given the effective base width – so it is necessary to calculate the p-region width of the emitter-base junction and the p-region width of the base-collector junction at equilibrium and subtract them both from the actual base width to find the effective base width.

\[
V_{BI} = \frac{kT}{q} \ln \frac{N_A N_D}{n_i^2} = 0.98 \text{ V}
\]

\[
W_{EB} = \sqrt{\frac{2\varepsilon_s}{q} \left( \frac{(N_A + N_D)}{N_A N_D} \right) V_{BI}} = 39.3 \mu\text{m}
\]

Solving for \(W_p\)

\[
N_A W_p = N_D W_n
\]

\[
W_n + W_p = W
\]

\[
W_p = \frac{W}{1 + N_A/N_D} \quad \text{So } W_{p(EB)} = 32.8 (\text{nm})
\]
This width will decrease slightly with a forward bias applied so let us use 30 nm.

Now for the base-collector region.

\[ V_{BL} = \frac{kT}{q} \ln \frac{N_{A}N_{D}}{n_{i}^{2}} = 0.92 \text{ V} \]

\[ W_{BC} = \sqrt{\frac{2\varepsilon_{s}}{q} \left( \frac{N_{A} + N_{D}}{N_{A}N_{D}} \right) V_{BL}} = 60.28 \text{ nm} \]

\[ W_{p(BC)} = \frac{W}{1 + \frac{N_{A}}{N_{D}}} = 20.1 \text{ nm} \]

So the effective base width \( W_{b} = 100\text{nm} - 30 \text{ nm} - 20.1 \text{ nm} = 50 \text{ nm} \)

\[ \beta = \frac{D_{nb}W_{e}N_{De}}{D_{pe}W_{b}N_{Ab}} \]

Gain \( \beta = \frac{1350}{480} \times \frac{1}{0.05} \times \frac{5 \times 10^{18}}{1 \times 10^{18}} \approx 280 \)

Question (continued): Estimate the gain of this transistor when a reverse bias of 8V is applied?

The width of the p-type region of the collector-base region under reverse bias of 8 V is calculated.

\[ W_{BC} = \sqrt{\frac{2\varepsilon_{s}}{q} \left( \frac{N_{A} + N_{D}}{N_{A}N_{D}} \right) (V_{BL} + V_{R})} = 187 \text{ nm} \]

\[ W_{p(BC)} = \frac{W}{1 + \frac{N_{A}}{N_{D}}} = 62.56 \text{ nm} \]

Then we take away the original 20 nm and get that the base width has decreased by 43 nm which gives an effective base width of 7 nm.

This gives a current gain of \( \beta = \frac{D_{nb}W_{e}N_{De}}{D_{pe}W_{b}N_{Ab}} = 2009 \)
THE EARLY EFFECT AND PUNCHTHROUGH

- As the collector/emitter voltage is increased, the reverse bias across the collector/base junction increases and the C/B depletion width therefore increases.
- This in turn leads to a narrower basewidth and hence an increased gradient of the electron distribution in the base.
- This increased gradient gives an increased collector current.
- This effect is known as the Early effect or basewidth modulation and gives an increase in collector current with collector/emitter voltage.

PUNCHTHROUGH

Punchthrough occurs when the neutral base width is reduced to zero at a sufficiently large $V_{BC}$. 