1. Diodes and Transistors

1.1 Introduction

The development of electronics over the last century and particularly in recent decades has been a great success story. Most of us probably own a mobile phone and make use of a computer at regular intervals throughout every day of our lives. In order to do this – we rely on electronic circuits. Even the ubiquitous car, washing machine and other every-day objects utilise electronic circuitry to enhance their performance.

In order to develop an understanding of how this electronic technology works, a good starting point is to study two of its building blocks – namely the diode and transistor. To begin this task, we will first examine the structure of the engineered semiconductor materials of which they consist. We will look at how charge is able to flow through these materials, or depending on the circumstances, how charge is prevented from flowing.

The following sections will introduce the notion of a 'band theory' of electron energies within various types of materials, before focussing on the semiconductor itself. The 'band theory' approach will be developed and set as a background to a description of how semiconductors are engineered into n-type and p-type materials in order that charge can be controlled and manipulated.

1.2 Band theory

Electrons in atoms are governed by rules which regulate their range of possible energy states. (For those that want to know more about this – read about the 'Pauli Exclusion Principle'). The consequence of this is that electrons in atoms are grouped together in discrete energy levels, each of which can only accommodate a limited amount of electrons. The lowest energy levels in an atom are always filled to

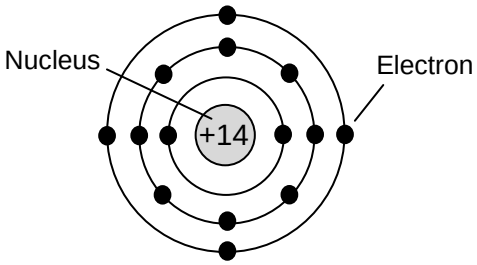

Figure 1.2.1: Structure of a silicon atom. The 14 electrons are grouped such that allowed energy levels, increasing from the centre outwards, are filled to capacity. Only the outer orbit has unoccupied energy states.

capacity before higher energy states can be occupied. The picture that emerges is for example, the atomic structure shown in Figure 1.2.1, where a silicon atomic nucleus is surrounded by filled electron energy levels or 'shells'.
The fact that each discrete energy level in an atom can only accommodate a limited number of electrons becomes a problem when we bring together many atoms of the same element (with identically filled energy levels) to form a solid. Here, the electrons begin to occupy the same space as electrons from adjacent atoms and consequently local energy levels are in danger of being overfilled. However, nature deals with this problem very neatly by creating energy 'bands' in solid matter which are formed by the 'splitting' of individual energy levels into many closely-spaced energy levels. Figure 1.2.2 demonstrates this by indicating how energy bands must widen as a function of decreasing distance between atoms to accommodate the new energy states. Just as there are no energy levels between the defined electron energies of an individual atom, so too are there 'forbidden' regions between the energy bands in a solid.

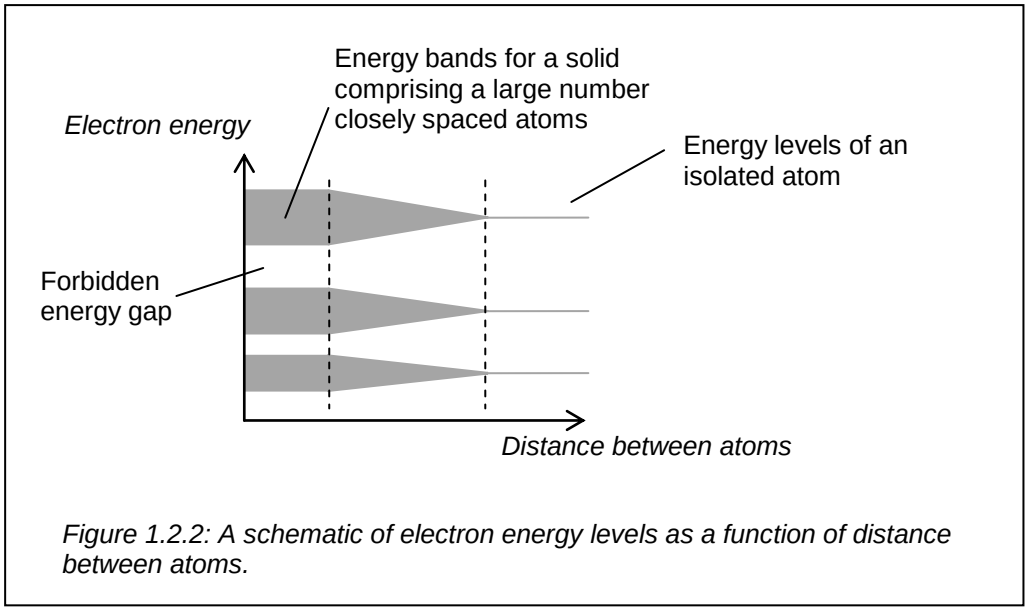

Figure 1.2.2: A schematic of electron energy levels as a function of distance between atoms.

Before we focus exclusively on the semiconductors used to produce electronic components, it is useful to first of all appreciate how the band structures of various type of material compare. Figure 1.2.3 shows the general band structure of three material types namely (a) conductor (b) insulator and (c) semiconductor. In each case, the upper two energy bands are shown: the valence band (enclosed within lines) and the conduction band (enclosed within dashes). The conduction band consists of electrons which are free to move within the atomic lattice whilst the valence electrons are typically bound to atoms, and help to form atomic bonds between atoms. In Figure 1.2.3, the shaded areas indicate filled energy levels.

In a conductor, it is typical to find that the conduction band overlaps the valence band, indicating a sharing of electrons and ease of mobility between these two bands. This is illustrated in Figure 1.2.3(a). However, it must be mentioned that in some examples of conductors, there is a small 'forbidden' gap between the valence band (filled to capacity in these cases) and conduction band. In either instance

nevertheless, the important thing to note is that the conduction band of a conductor is always *partially* filled with electrons. The part-filled conduction band allows electrons to move to a slightly higher energy level which they must do in order to be

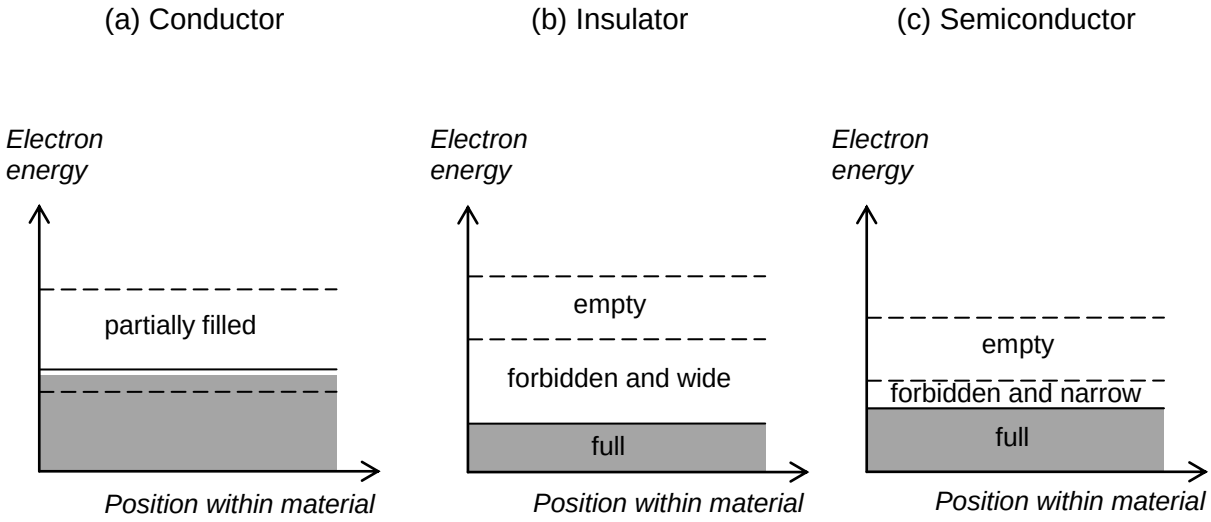

Figure 1.2.3: Upper two energy bands of a (a) conductor, (b) insulator and (c) semiconductor. All lower bands (not shown) are filled. The valence band (lines) and conduction bands (dashes) are shown.

mobile when a current is produced. (An electron that is moving has increased energy over one that is not moving). The flow of current is therefore possible in a conductor because of the many vacant energy levels in the conduction band.

On the other hand, the insulator shown in Figure 1.2.3(b) is characterised by a very large gap between the valence and conduction bands which is too wide for an electron to cross. Moreover, the valence band is filled to capacity. Therefore there are no vacant states to which electrons can move in an insulator and consequently current cannot flow.

Finally, Figure 1.2.3(c) shows the band structure of a semiconductor. Just like the insulator its valence band is full and its conduction band empty. However, here the band gap is narrow enough for the occasional electron to gain enough thermal energy at room temperature to jump across from the highest level in the valence band to the lowest level of the conduction band. As a consequence, the semiconductor conduction band is not always *completely* empty. The thermal transition of electrons forms the basis of why the semiconductor is most suitable for manufacturing electronic components where, after careful engineering of the semiconductor material, a high degree of control over the flow of electrons is possible. However, it also indicates why semiconductor electronics can be temperature-sensitive.

1.3 The doping of semiconductors

Before we examine how we engineer the structure of a semiconductor to make it suitable for the manufacture of an electronic component, let's first consider what happens to the band structure of a semiconductor when an electron is promoted from the valence band to the conduction band by thermal energy.

In Figure 1.3.1, the valence and conduction bands of a piece of semiconductor material are again shown, this time with a voltage connected across it consisting negative and positive terminals. Let's now assume that an electron has been thermally promoted to the conduction band. This electron will now migrate towards the positive potential resulting in a (small) current through the conduction band. But – now there is a 'hole' left in the valence band formed by the electron's previous energy state. Another electron in the valence band will now be shaken loose from its position by the influence of the positive potential and will be promoted upwards and to the right to fill the vacant state. The process continues with a further electron being attracted towards the positive potential, and upwards in energy, to the position of the previous electron and so on. We can think of the 'hole', which is migrating ever closer to the negative potential, as being a positively-charged particle producing a 'positive-particle' current.

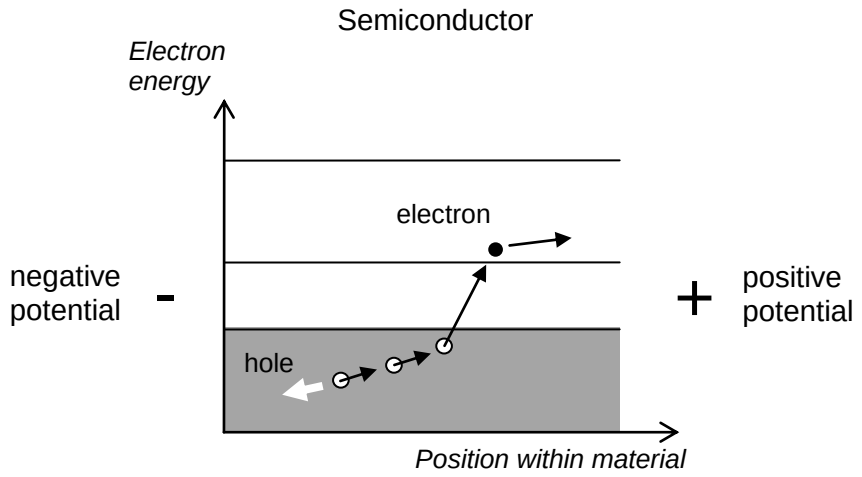

Figure 1.3.1: A negatively-charged electron is thermally excited from the valence band to the conduction band leaving behind a vacant energy state or 'hole'. The positively charged 'hole' will then migrate towards the negative potential through the valence band as nearby electrons fill the hole by being attracted to the more positive potential.

The overall current produced by thermal excitement of electrons in a pure semiconductor, known as an *intrinsic semiconductor*, is very small. However, it is possible to increase the level of charge carriers in a controlled manner by 'doping' the pure semiconductor with impurities. A semiconductor thus doped is called an *extrinsic semiconductor*. The crystal lattice of the semiconductor can be doped in

order to produce either an excess of electrons (known as n-type doping) or holes (known as p-type doping).
Figures 1.3.2 and 1.3.3 below demonstrate the effects of n-type and p-type doping by showing the ensuing energy band diagrams and a section of the semiconductor crystal lattice in each case.

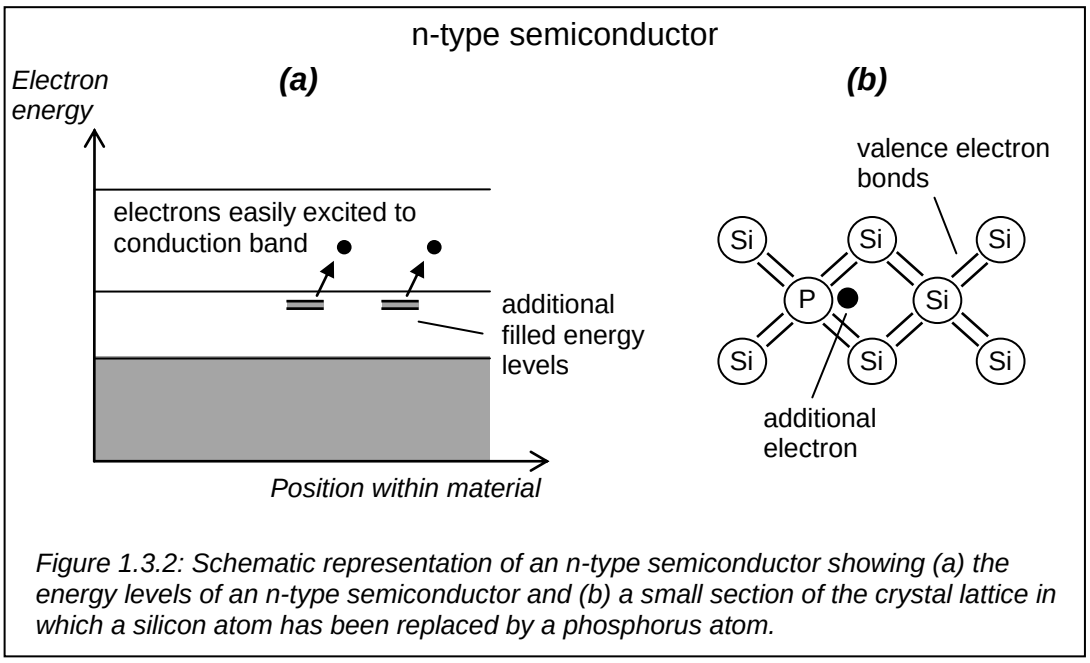

Figure 1.3.2: Schematic representation of an n-type semiconductor showing (a) the energy levels of an n-type semiconductor and (b) a small section of the crystal lattice in which a silicon atom has been replaced by a phosphorus atom.

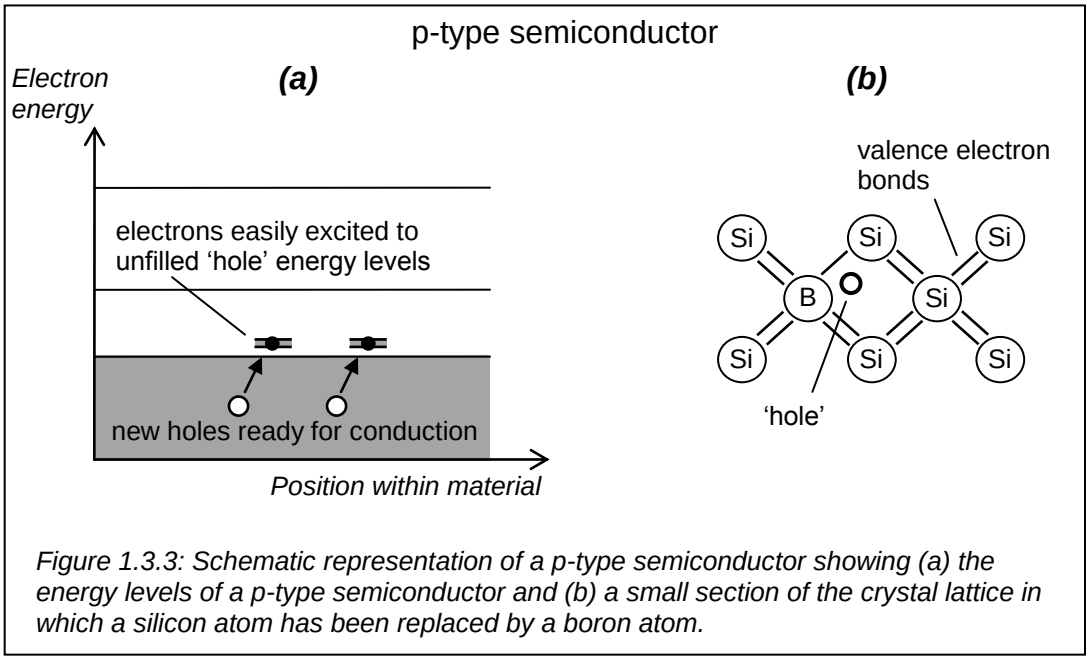

Figure 1.3.3: Schematic representation of a p-type semiconductor showing (a) the energy levels of a p-type semiconductor and (b) a small section of the crystal lattice in which a silicon atom has been replaced by a boron atom.

Let us consider that we choose to 'dope' silicon (Si), which is by far the most abundantly utilised semiconductor in electronics. The silicon atom has an outer valance shell with four electrons as we saw in Figure 1.2.1, a few pages back. The four valence electrons form covalent bonds with other neighbouring silicon atoms in

order to create a crystal silicon lattice. Each silicon atom actually bonds with the lattice via eight electrons, four of its own, and one from each of the four neighbouring atoms, indicated by the eight bonding lines on one of the silicon atoms in Figures 1.3.2(b) and 1.3.3(b).

In doping the silicon, impurities from Group V such as phosphorus (P) or from Group III such as boron (B) can be used to produce n-type and p-type semiconductors respectively.

When intrinsic silicon is doped with phosphorus for example, the phosphorus atoms are found to occupy atomic sites normally occupied by silicon atoms. However, given that phosphorus atoms have five valence electrons and only four are needed for bonding, there remains one loosely bound electron per phosphorus atom. This is indicated in Figure 1.3.2(b). The loosely bound electrons have an energy level that is just below the conduction band of the overall material and as a consequence are easily excited to the conduction band at room temperature, as shown in Figure 1.3.2(a). The Group V impurities, known as *donor impurities*, make it easier for the semiconductor to conduct. This resulting doped semiconductor is called n-type because the charge carriers are (n)egatively-charged electrons.

If the silicon is doped with boron, once again the boron atoms are found to occupy sites normally occupied by silicon atoms. However, given that boron has only three valence electrons, there will be a 'hole' in the crystal lattice associated with each boron atom where a bonding electron is missing, as shown in Figure 1.3.3(b). The energy levels of the holes are such that they are just above the valence band and are therefore easily filled at room temperature by nearby valence electrons. Although electrons that fill the vacant energy levels are not able to contribute to the current, doping with these Group III *acceptor impurities* cause there to be an abundance of holes in the valence band which *can* flow through the material. This type of doped semiconductor is called p-type because the charge carriers are (p)ositively-charged holes.

In the n-type material, the electrons are called *majority charge carriers* and holes (caused by random thermal excitations) are called *minority charge carriers*. Similarly, in the p-type material, the holes are called *majority charge carriers* and the electrons (again excited from valence to conduction band by random thermal excitation) are called *minority charge carriers*.

1.4 The p-n junction (diode)

Our first circuit component, the semiconductor diode, is basically an n-type semiconductor 'joined' to a p-type semiconductor. In reality, the diode (or p-n junction) is manufactured from a *single* crystal of semiconductor material, but it is instructive to consider what would happen if we gradually brought the two materials together to form the diode, which is the approach we take here. Figures 1.4.1(a) and 1.4.1(b) illustrate the band structure of the two materials before they are brought together. The doping has caused there to be an excess of electrons in the conduction band of the n-type material and an excess of holes in the p-type material

as described in the previous section. The small number of holes in the n-type material and small number of electrons in the p-type material are caused by random thermal excitations from valence to conduction bands. It must be noted that in both materials the overall charge distribution is initially *neutral*.

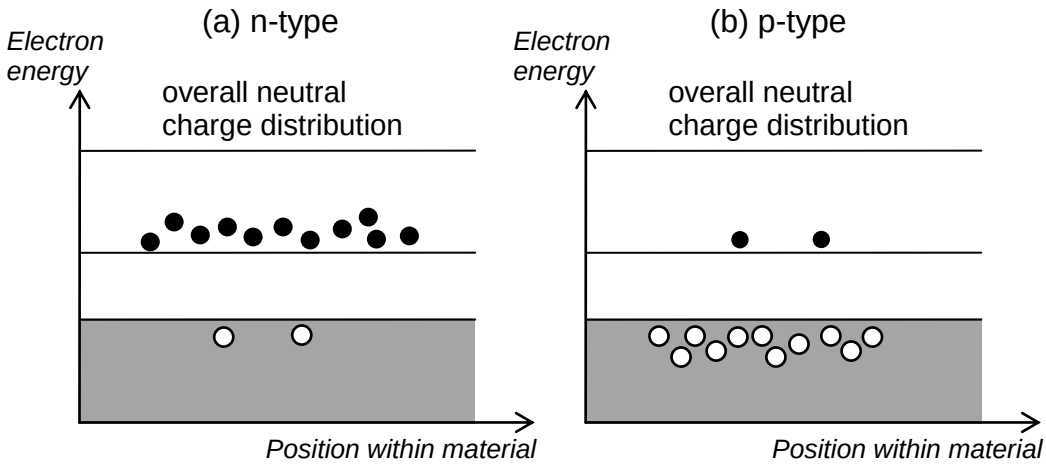

Fig 1.4.1: The n-type and p-type semiconductors before being 'brought together' to form a diode

If we now bring the two materials together into one slab to form the p-n junction, initially the electrons near the boundary will diffuse from the higher density conduction band of n-type material to the lower density conduction band of the p-type material. Similarly, the higher density of holes near the boundary in the p-type material will begin to diffuse into the n-type material. (Essentially in both cases, electrons are diffusing from the n-type to the p-type material). *Now there is no longer charge neutrality:* the p-type side of the junction is more negatively charged having gained electrons, and the n-type side of the junction becomes more positively charged having lost electrons. Consequently, an electric field *E*, pointing from the n-type to p-type material is created.

The electric field at the boundary will grow in magnitude until further net electron migration is prevented: a balance is struck between electrons migrating across the boundary in one direction (n-type to p-type) as a consequence of diffusion, and drifting in the other direction (p-type to n-type) as a consequence of the electric field. This is clearly dependent upon temperature also, given that the final equilibrium condition can be affected by the population of thermally excited electrons (from valence to conduction bands), which has an effect on the rate of diffusion.

A complicated picture emerges where holes on both sides of the boundary become (mostly) filled by electrons originating predominantly from the conduction band. In the case of silicon, electron energy lost by the 'jump' from conduction to valence band is given up to the crystal as vibrational energy. (Read about 'phonons' which are units of quantum-mechanical vibrational motion if you are interested in this sort of detail). Very occasionally the energy is given up as a photon (a particle of light) as a consequence of the energy jump. (This is the basis of the light-emitting diode or

LED, although silicon is not a great material for giving up photons in this manner – again, read about 'direct' and 'non-direct' band-gap materials if you are interested in why some semiconductor materials are better than others for generating light).

The situation that emerges can be seen in Figure 1.4.2 where a thermally 'balanced' p-n junction is shown with both its band diagram (a), and physical structure (b).

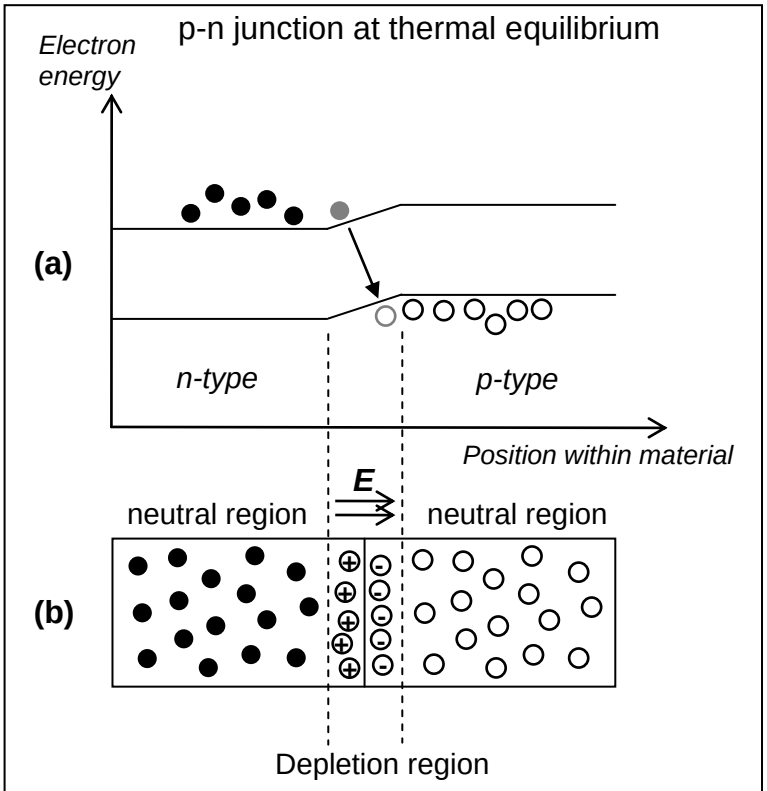

Fig 1.4.2: The p-n junction (diode) at thermal equilibrium. (a) The band diagram and (b) the physical structure

Inspection of the band diagram in Figure 1.4.2(a) shows the energy bands in the p-type material now slightly raised above those in the n-type material. This is a consequence of the Electric field *E* also shown in the Figure. Negatively-charged electrons that attempt to migrate across the boundary from n-type to p-type conduction bands will need to swim 'uphill' and therefore need to migrate to significantly higher energy levels if they are not already in a higher energy state. (They are basically being repelled by the negative charge that has accumulated on the p-type side of the boundary and therefore only the most energetic will successfully cross to the p-type conduction band). Those electrons that ultimately combine with holes as can be seen in Figure 1.4.2(a) subsequently give up some of their energy after crossing the boundary as described above.

Figure 1.4.2(b) shows the physical structure of the diode. The main bulk of the diode remains neutrally charged (i.e. there are the correct number of electrons for all the atoms in the structure). However, there is a region indicated by the vertical dashes called the *depletion region* which spans a small distance either side of the boundary.

The width of the depletion region is fixed once a thermal-equilibrium condition is reached because there is no longer a net migration of electrons. On the n-type side of the depletion region there are electrons missing so positive ions exist. On the p-type side of the depletion region electrons fill holes to produce negative ions. The depletion region becomes an important element in the discussions that follow regarding diodes because of its 'built-in' potential difference.

Finally, before we begin to study in the following section how the diode can be utilised in a circuit, let's observe what happens if we take our manufactured, thermally stable p-n junction and apply a potential difference across its terminals. With a potential difference first applied one way, then the opposite way, Figure 1.4.3 demonstrates the defining characteristic of a diode: *the diode has a small resistance to current flow in one direction (forward bias) and a very high resistance to current flow in the other direction (reverse bias)*.

Figure 1.4.3(a) shows our semiconductor diode in reverse bias, with the positive terminal of the supply voltage, *V* connected to the n-type material. As can be seen,

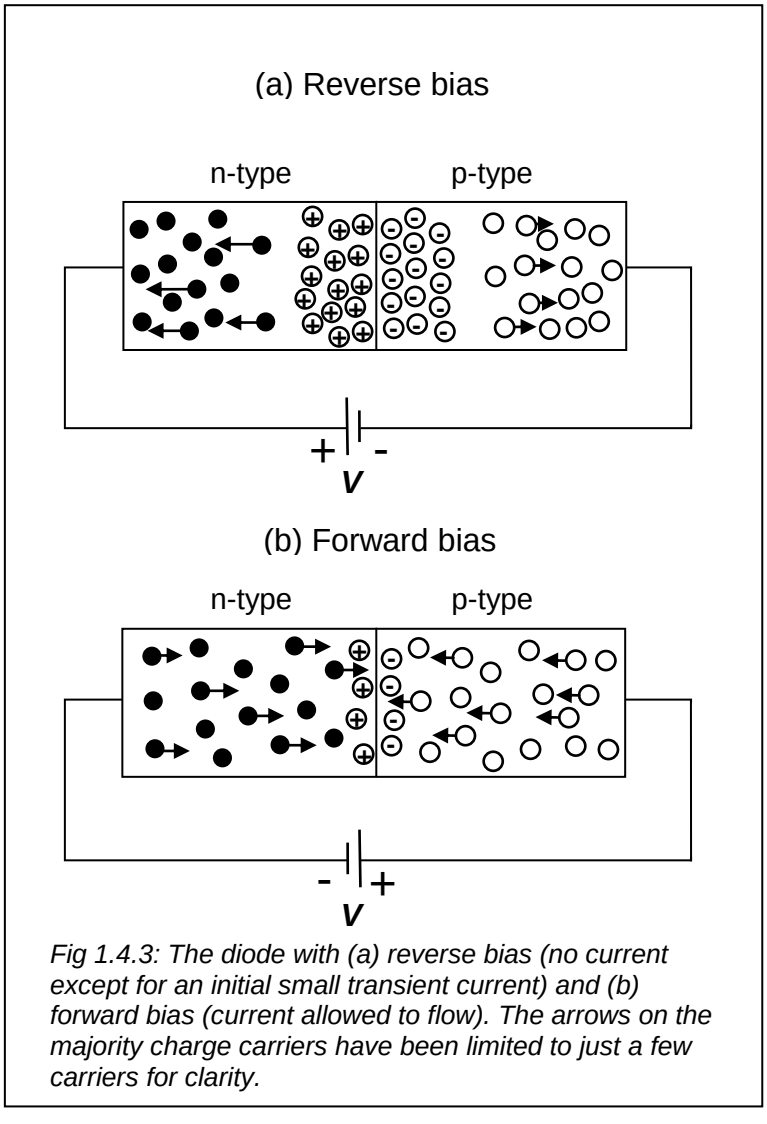

Fig 1.4.3: The diode with (a) reverse bias (no current except for an initial small transient current) and (b) forward bias (current allowed to flow). The arrows on the majority charge carriers have been limited to just a few carriers for clarity.

the negatively-charged free electrons in the n-type material are attracted towards the more positive terminal of the supply, and the positively-charged holes in the p-type material are attracted towards the more negative terminal of the supply. Consequently, after a small transient movement of charge, current cannot flow due to the separation of electrons and holes. (However, it has to be mentioned that the current isn't exactly zero because of thermal effects. As mentioned several times already, the occasional electron will always find a way to be excited thermally from valence to conduction bands. If this happens within the depletion region, the conduction electron will be swept along by the electric field. This is known as *reverse leakage current*). Note that in reverse bias, the depletion region has widened.

Figure 1.4.3(b) on the other hand, shows our semiconductor diode in forward bias, with a supply voltage of opposite polarity. This time our free electrons drift in the opposite direction towards the more positive supply, and the holes drift towards the more negative supply. (You could also think of the electrons being repelled by the negative supply and the holes being repelled by the positive supply – it's equivalent). Holes and electrons meet in the depletion region whereupon electrons from the n-type material 'drop' into the holes (as was shown in Figure 1.4.2) and continue drifting across the p-type material (or in other words, the holes in the p-type move towards the n-type material). Current is now allowed to flow. As can be seen, the depletion region still exists but is less wide than if the supply is either reversed or not connected at all.

One question still remains – what about the electric field that exists across the depletion region that, given its direction, ought to oppose the flow of current in forward bias? Indeed there is a 'step' in the potential across a diode. In the case of a silicon diode in forward bias, this is about 0.7 Volts and for a germanium diode about 0.3 Volts. (Diodes are mostly fabricated with silicon although germanium is also sometimes used.) This means that for a forward-biased silicon diode, our supply voltage needs to be at least about 0.7 Volts in order to produce current flow.

1.5 The semiconductor diode as a circuit element

The forward-biased diode is shown once again in Figure 1.5 in order to introduce the circuit symbol for the diode, but also to make clear the symbol's orientation relative to the p-n junction. The 'arrow' on the circuit symbol always points from p-type to n-type, and in the case of forward-bias, in the direction of *conventional* current flow. The p-type side is referred to as the 'anode' and the n-type side is referred to as the 'cathode'.

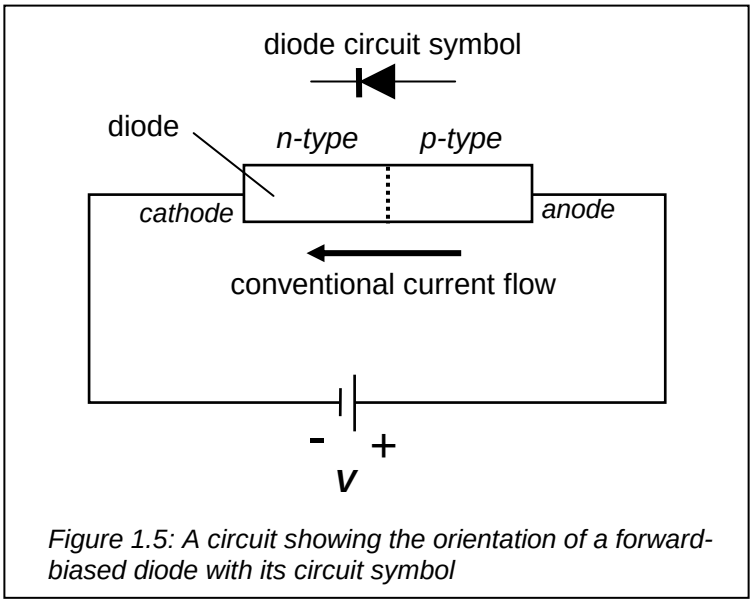

Figure 1.5: A circuit showing the orientation of a forward-biased diode with its circuit symbol

1.5.1 Diode voltage-current (V-I) characteristics

The circuit shown in Figure 1.5 can be adapted to record the *voltage-current* characteristic of a diode by inserting an ammeter of suitable sensitivity into the

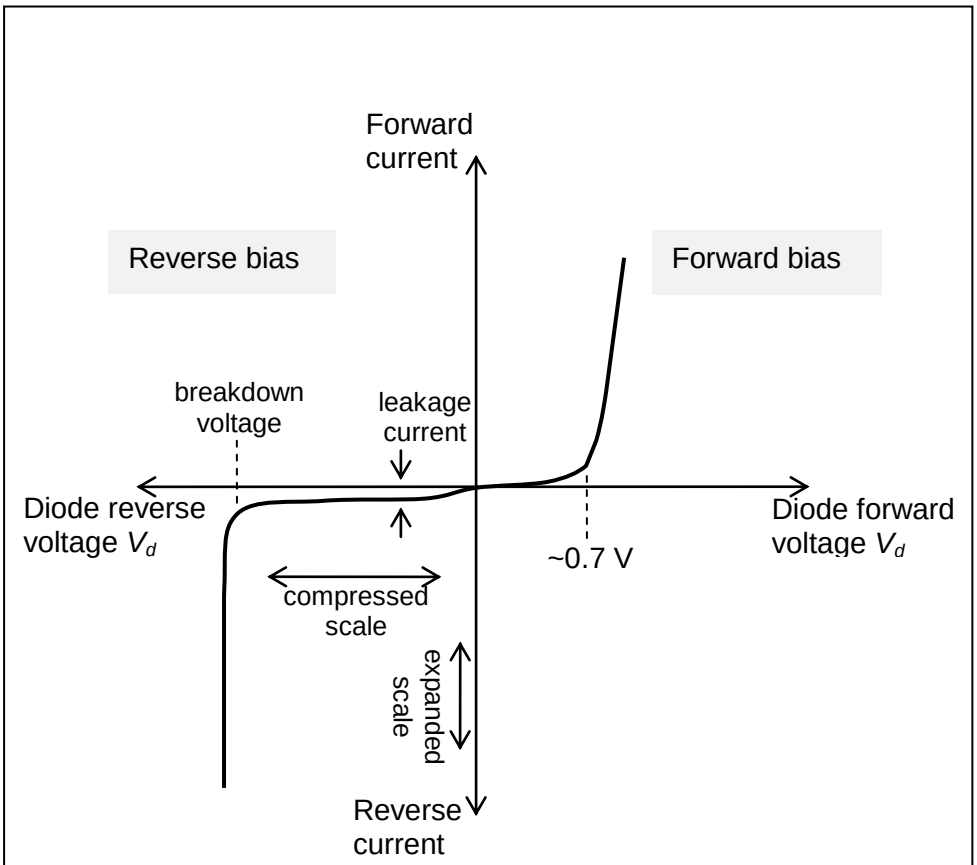

Figure 1.5.1: Silicon diode voltage-current (V-I) characteristic. Note the respective 'compressed' and 'expanded' voltage and current scales of reverse bias relative to forward bias.

circuit to record the circuit current (and thus the diode current) whilst connecting a voltmeter across the terminals of the diode to record the diode voltage V_d. Typical *voltage-current* characteristics of a silicon diode are shown in Figure 1.5.1. Although for most applications we can consider a semiconductor diode to be essentially closed-circuit in forward bias and open-circuit in reverse bias, as can be seen, the overall picture is not so straightforward. As explained in the previous section, in forward bias the voltage needs to be at least about 0.7 V to allow current to begin to flow normally. The voltage will increase very little beyond this value even though the circuit current may vary. (You may find that ~0.6 V or other voltage is quoted elsewhere for V_d at the point where current begins to flow. As can be seen there is a non-zero gradient at this point and beyond. Various texts make use of different definitions).

In summary, we can say that when the silicon diode is forward biased and current is flowing, the voltage drop across the diode is approximately 0.7 V.

In Figure 1.5.1, the full-scale of the axis representing the forward current is of the order tens of milliamps (mA) in contrast to the expanded full-scale of the axis representing the *reverse* current, which is of the order tens of microamps (µA). The scaling allows the 'reverse leakage current' mentioned in the previous section to be observed. The leakage current generally has a magnitude of a few microamps or less, although this clearly depends on temperature.

Conversely, the diode reverse voltage axis has been *compressed*. As can be seen, at a particular reverse voltage, which can be hundreds of volts (depending on design and doping levels), there is a particular 'breakdown voltage' beyond which current is allowed to flow normally. The steepness of the graph at this point leads us to conclude that the voltage across the diode remains fixed whilst current can vary. The phenomena can be used to advantage and is discussed in section 1.5.4 below in the context of 'Zener diodes'.

1.5.2 Diode Rectification

A widespread use for diodes is *rectification*. This is a process which converts AC into DC and is particularly important for electronic circuits, given that these circuits normally operate on low voltage and with direct current. Here, we will not study how we first convert higher AC mains voltages to lower AC voltages, which is necessary *before* rectification. This process of amplitude reduction is achieved with a *transformer*, which you can find out about if you would like a more complete picture about how DC power supplies are made. Here, we will assume that the amplitude of the AC source is sufficiently lowered to suit a given purpose in order that rectification can be performed directly.

1.5.2.1 Half-wave rectification

Figure 1.5.2.1(a) demonstrates the simplest form of rectification, *half-wave rectification*, with a series circuit that consists of an AC supply, a diode and a load

resistor. The voltage across the resistor forms the output. The voltage/time graph in Figure 1.5.2.1(b) shows how the AC input voltage v_{in} (black curve) compares to the DC output voltage across the resistor v_L (red curve). As can be seen, this is not a very efficient arrangement because half of the signal is lost between input and output. This is because the diode is allowing current to travel in one direction only (hence half-wave rectification). Moreover, the DC output signal is very 'bumpy' and therefore not desirable for many applications. Over the next few sections we will see how these parameters can be improved. Note that the output curve has a reduced peak value compared to the input curve because as discussed before, there is a voltage drop of about 0.7 V across the diode.

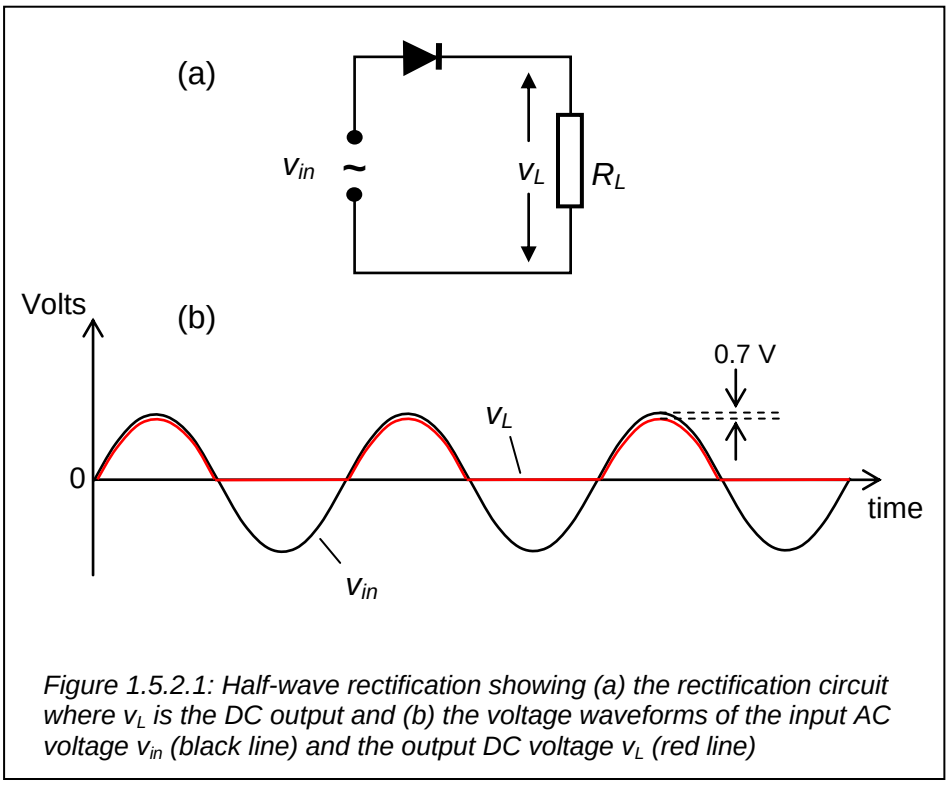

Figure 1.5.2.1: Half-wave rectification showing (a) the rectification circuit where v_L is the DC output and (b) the voltage waveforms of the input AC voltage v_{in} (black line) and the output DC voltage v_L (red line)

1.5.2.2 Full-wave rectification

We can improve the low efficiency observed in half-wave rectification by building a circuit which includes four diodes resulting in *full-wave rectification*. The circuit is shown in Figure 1.5.2.2(a), again with v_{in} as our AC input and the voltage v_L across R_L forming the DC output. The direction of current around the circuit is indicated as i_{+ve} when the AC input voltage is in the forward direction and i_{-ve} when the AC input voltage is in the reverse direction. The current now flows through the load resistor in a single direction whether the input voltage is positive *or* negative. As can be seen in the graph of Figure 1.5.2.2(b), there is now a forward–biased DC voltage across the load resistor over the duration of each full cycle (hence full-wave rectification), despite the voltage at the input changing direction at each half-cycle. There are several things to note with this circuit. Firstly, given that the current passes through two diodes instead of one, we now have a voltage drop in the peak value from input to output of 2 x 0.7V = 1.4V. Secondly, the frequency of the output signal is now twice that of the input signal. Thirdly, although the circuit has improved efficiency

over half-wave rectification, the output signal is still rather 'bumpy'. We will see what can be done to 'smooth' the signal in the next section.

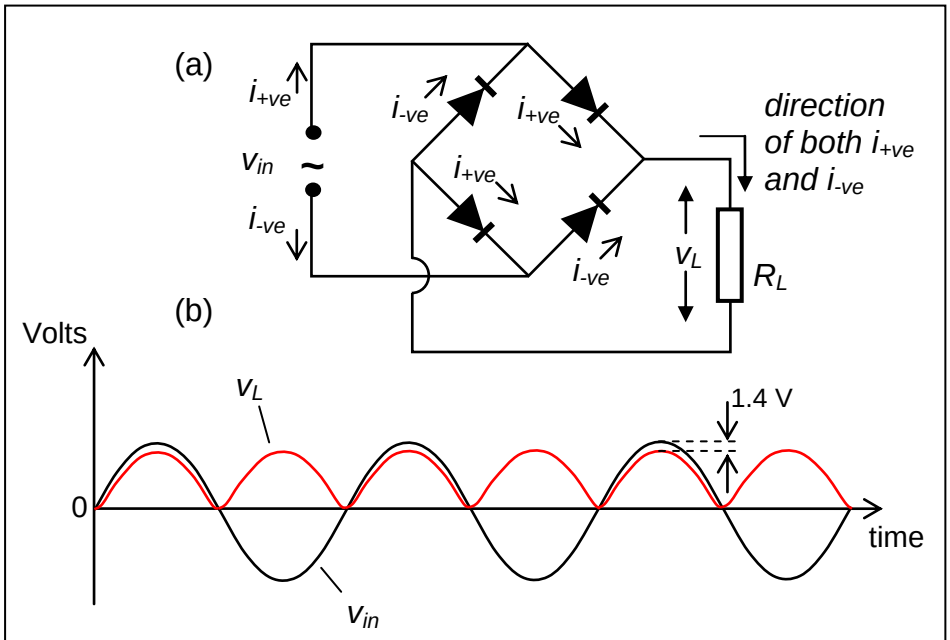

Figure 1.5.2.2: Full-wave rectification showing (a) the rectification circuit where v_L is the dc output and (b) the voltage waveforms of the input ac voltage v_{in} (black line) and the output dc voltage v_L (red line)

1.5.3 Smoothing and ripple reduction

The bumps observed in the output signal of the full-wave rectifier can be 'smoothed'

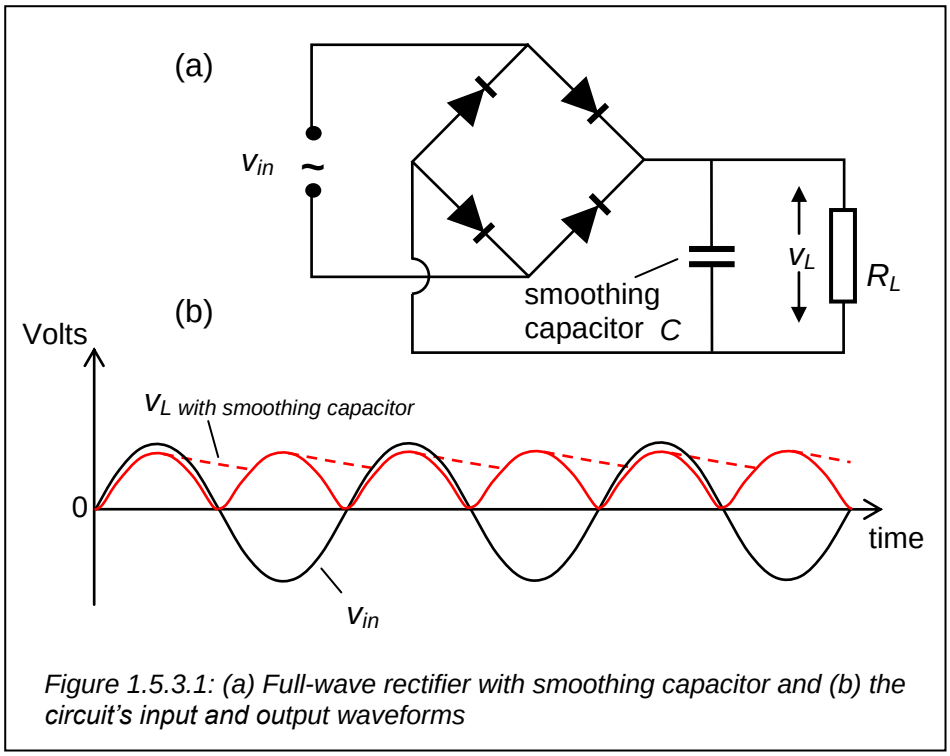

Figure 1.5.3.1: (a) Full-wave rectifier with smoothing capacitor and (b) the circuit's input and output waveforms

in order to produce a signal approaching a constant voltage supply. Figure 1.5.3.1(a) shows how the full-wave rectification circuit can be adapted to achieve this. The simple addition of a capacitor across the output of the rectifier has the effect of smoothing the output as shown graphically in Figure 1.5.3.1(b). The resulting *moderate* undulations in the signal are collectively known as 'ripple'. A smoothing capacitor added to the circuit with a well-chosen capacitance will charge up to the peak value of the output voltage before discharging slowly through the load resistor R_L as the input signal falls in magnitude more rapidly. In this way, the voltage across the load resistor is maintained at a higher level for longer. The orientation of the diodes ensures that the only path for capacitor discharge is through the load resistor. If we make the time constant R_LC large compared to the period of the rectified signal, the capacitor voltage will fall only marginally before it is recharged during the following cycle. (For more information about the rate of discharge and indeed rate of charge of capacitors – read about *DC transients*. Essentially, a capacitor will take roughly 5 time constants to fully discharge with an exponentially-decaying function. The time constant of discharge is equal to the capacitor's capacitance in Farads C, multiplied by the total resistance to the discharging current, in this case R_L, i.e. the time constant here, $T = R_LC$).

It is possible to calculate by how much the voltage ripple will fall and rise for a given value of capacitance. By the same token, by knowing what level of ripple is acceptable for a particular application, we can choose an appropriate smoothing capacitor for the circuit. The following analysis derives useful equations for this purpose. For clarity, Figure 1.5.3.2 shows the output ripple seen in 1.5.3.1(b) and defines two parameters:

(i) ΔV – the fall in voltage from the peak voltage and
(ii) Δt – the period of the rectified signal

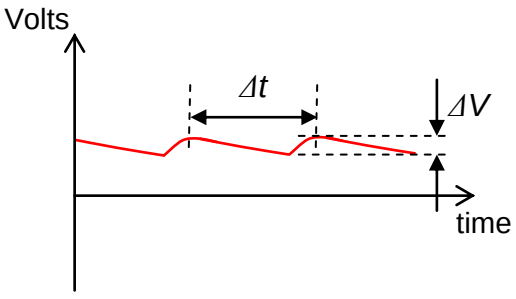

Figure 1.5.3.2: Defining ΔV and Δt within the ripple of the full-wave rectified signal

We have seen earlier in the module that in the case of capacitors the following relationship applies:

$$Q = CV \qquad (1.1)$$

where a charge of magnitude Q will be stored on a capacitor of capacitance C with an applied voltage of V across its plates. If we take the time derivative of both sides we get

$$\frac{dQ}{dt} = C\frac{dV}{dt} \tag{1.2}$$

The left-hand-side of Equation (1.2) now represents the flow of charge in the vicinity of the capacitor i_{DC}. (We will see that we intend to equate the capacitor current to the DC load current later in the derivation hence the subscript in i_{DC}). We can therefore write

$$i_{DC} = C\frac{dV}{dt} \tag{1.3}$$

However, we could say that the ripple effect seen in Figure 1.5.3.2 is approximately triangular in shape displaying an approximately linear fall in voltage followed by a steep rise in voltage over a short time-scale. (These approximations are valid for small ripple magnitude). We could therefore approximate Equation (1.3) as

$$i_{DC} = C\frac{\Delta V}{\Delta t} \tag{1.4}$$

where i_{DC} can now be taken to be the *average* current through the load resistor that we are seeking to maintain with the smoothing capacitor.

Since Δt is equivalent to half of the AC period T and $T = 1/f$ we can write

$$i_{DC} = C\frac{\Delta V}{\Delta t} = C\frac{2\Delta V}{T} = 2Cf\Delta V \tag{1.5}$$

therefore

$$\boxed{C = \frac{i_{DC}}{2f\Delta V}} \tag{1.6}$$

If we subtract $\Delta V/2$ from the peak voltage V_p we will acquire the average output voltage thus

$$\boxed{V_{average} = V_p - \frac{i_{DC}}{4fC}} \tag{1.7}$$

NOTE: When designing a full-wave rectification circuit with a particular DC output voltage, it is important to take account of the voltage drop across the diodes. Moreover, the 'target' DC output voltage will be the *average* voltage. For example if the voltage required is 5 V with a ripple ΔV = 0.2 V, then the design will constitute 5 ± 0.1 V. Finally, it is important to recognise that if a smoothing capacitor is used on a half-wave rectifier, then the capacitor value will increase by a factor of two because the rectified signal will have a frequency equivalent to the AC input frequency i.e. $\Delta t = T$.

1.5.4 Zener diodes

Section 1.5.1 and Figure 1.5.1 shown earlier described some voltage-current (V-I) characteristics of a diode, particularly when forward biased. However, a more complete description of the characteristics of reverse bias has been left for this section in order to introduce a special kind of diode known as a *zener diode*.

If a diode is reverse biased, as shown in Figure 1.4.3(a), electrons and holes are attracted away from the depletion region preventing current from flowing. This leads to a widening of the depletion region as previously noted and consequently an increase in the potential difference across the region. The increase in the potential difference across the region necessarily means a greater electric field across the region too. If the reverse bias is large enough, the depletion region will acquire an electric field of sufficient magnitude to break the covalent bonds of its atoms. This 'tearing away' of the electrons from their atoms produces a large number of free electrons and holes which rapidly increase the reverse current.

The breakdown of the atoms in this way is known as *zener breakdown* and occurs at the *breakdown voltage* indicated in Figure 1.5.1. (It has to be noted that there is another mechanism for breakdown known as 'avalanche' breakdown which we will not discuss here). The zener breakdown voltage can be anywhere from about fifty volts (for a typical rectifier diode) to hundreds of volts and beyond for some specialised diodes.

Zener breakdown can be used to advantage. By heavily doping a diode, breakdown can occur when a moderate reverse voltage is applied to the diode. In fact a diode can be engineered such that we know precisely at what voltage it will undergo breakdown, allowing current to flow. Diodes produced in this way are called *zener diodes*. Given that the (V-I) characteristic at breakdown (as shown in to Figure 1.5.1) is very steep we can conclude that when current flows in a reverse direction through a zener diode, the zener (breakdown) voltage remains (to a very good approximation) constant across the diode, even if there are fluctuations in the voltage supply of the circuit containing the zener diode, or the magnitude of current passing through the diode varies.

This property informs us that *zener diodes are useful as voltage regulators* as we will see in the following section.

The circuit symbol for a zener diode is introduced in Figure 1.5.4 where its orientation is shown relative to (the backward-flowing) conventional current.

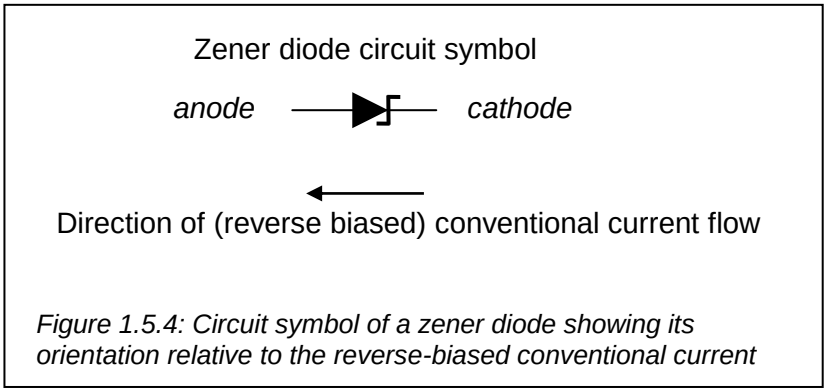

Figure 1.5.4: Circuit symbol of a zener diode showing its orientation relative to the reverse-biased conventional current

1.5.5 Regulation

The constant voltage produced across a zener diode when a reverse current is flowing through it can be used to advantage when voltage needs to be regulated, i.e. held steady, particularly under a variable load, or ripples in the supply. A simple zener diode regulator circuit is shown below in Figure 1.5.5.

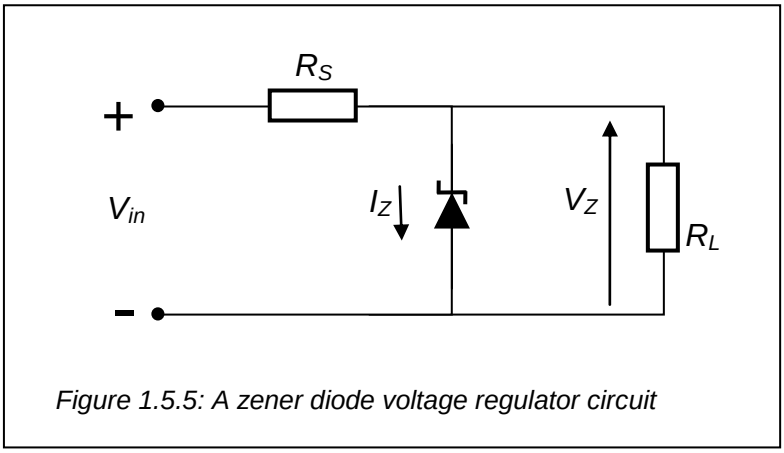

Figure 1.5.5: A zener diode voltage regulator circuit

As can be seen in Figure 1.5.5 above, the zener diode is positioned in parallel to the load in order to regulate the load voltage, V_Z. The resistor R_S in series with the diode limits the current passing through the diode I_Z. The current flow through the diode is limited because if excessively high, it will cause the device to be damaged. (Manufacturers always make known the individual power ratings of their zener diodes for this reason).

Given a particular load current and required diode current (normally about 5mA), the value of R_S is straightforward to calculate:

$$R_S = \frac{V_{in} - V_Z}{I_{load} + I_Z} \qquad (1.8)$$

or in other words, the series resistor value equals the required voltage across the resistor divided by the current through it (from V=IR).

Zener diodes are available commercially with given voltages V_z ranging from about 1.8V to 200V.

1.5.6 Diode limiter and clamp circuits

In this section we conclude our discussion about diodes by returning to circuits that consist of conventional diodes, i.e. those that are designed to be used in forward bias.

Sometimes it is necessary to introduce voltage protection circuitry to protect sensitive electronic circuits from spikes or fluctuations in the supply. Protection circuits of this kind are called *limiters*, and are used to ensure that an output voltage never exceeds a given level. Three examples of limiter circuits are shown in Figures 1.5.6.1(a), (c) and (e), with their respective voltage outputs plotted against their voltage inputs in Figures 1.5.6.1(b), (d) and (f).

In Figure 1.5.6.1(a) the diode will only allow current to pass when it is forward biased, therefore when V_{in} is positive, the circuit will function as if the diode was not there. However, if V_{in} becomes negative, or more precisely a negative value less than about -0.7V, then the diode will be forward biased and will hold V_{out} at a potential difference of (about) -0.7V for all values of $V_{in} < -0.7V$. This is shown in the graph in Figure 1.5.6.1(b).

Figure 1.5.6.1(c) shows a similar circuit, except this time it includes two diodes of opposing direction. Here, one or other of the diodes will conduct if V_{in} has either a value greater than 0.7V or less than -0.7V. Hence V_{out} is bounded by these two voltages as can be seen in Figure 1.5.6.1(d).

Figures 1.5.6.1(e) and (f) show how we can vary the threshold of the limiter output voltage by the inclusion of a battery to the circuit. In this example, V_{in} must exceed the value $V_x + 0.7V$ in order for the diode to conduct and hence limit the output. This introduces flexibility into the circuit for different applications.

Figure 1.5.6.1(g) introduces a different kind of circuit called a *clamp circuit* which is used to shift an AC signal by a constant voltage. In the example given, when the input voltage becomes less than -0.7V, the capacitor will begin to charge with a negative and positive potential on its plates as indicated. This is because under this condition, the diode will conduct allowing current to flow, and thus allowing the capacitor to charge up. (The negative supply will repel electrons which collect on the left-hand-side capacitor plate, thus repelling electrons from the right-hand side capacitor plate which pass through the diode towards the more positive potential of the supply. This description is valid even if the ‘more positive’ terminal is fixed at 0V).

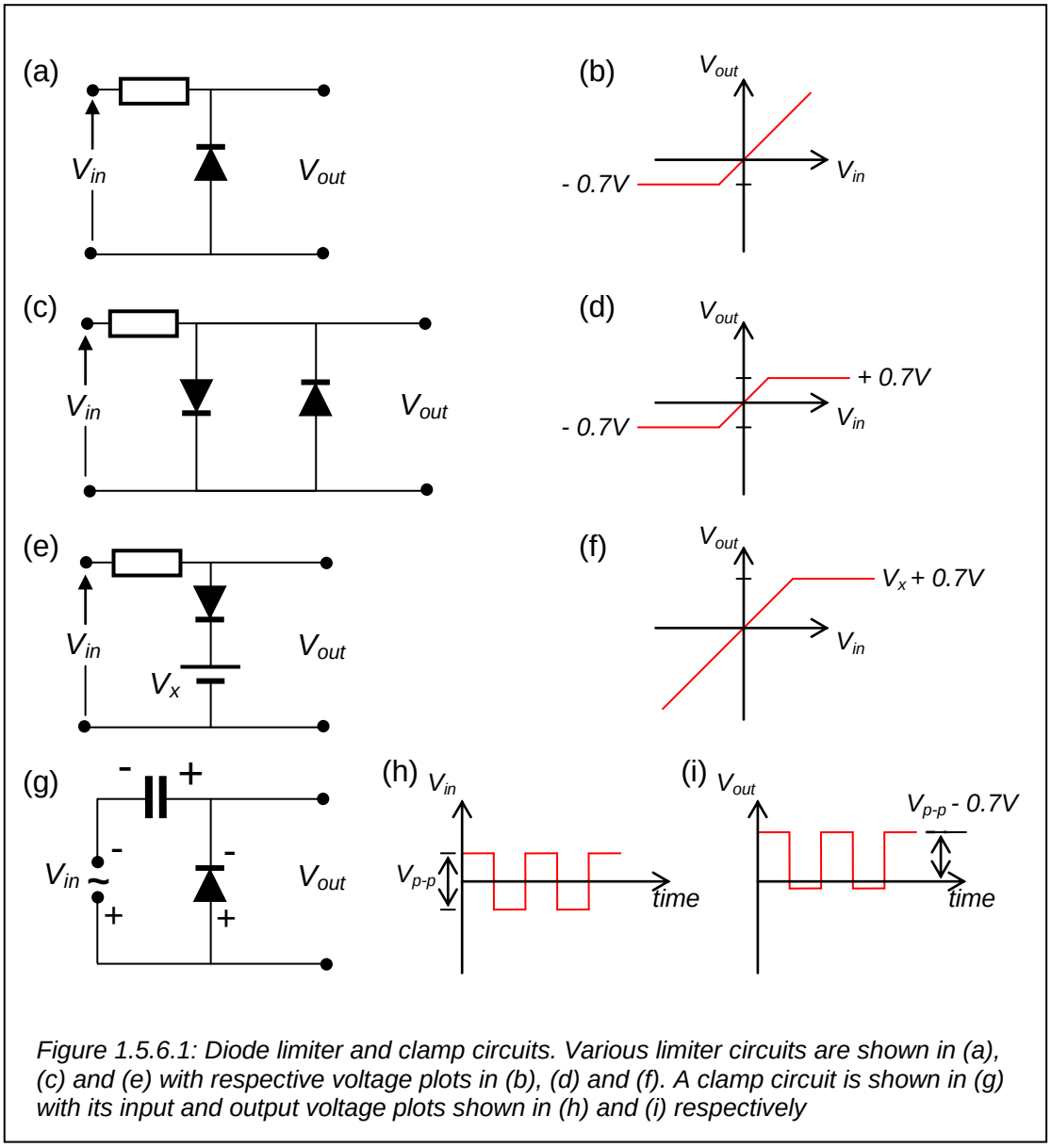

Figure 1.5.6.1: Diode limiter and clamp circuits. Various limiter circuits are shown in (a), (c) and (e) with respective voltage plots in (b), (d) and (f). A clamp circuit is shown in (g) with its input and output voltage plots shown in (h) and (i) respectively

As the AC cycle continues, the capacitor will eventually charge up to a potential difference of V_p - 0.7V where V_p is the peak voltage (or amplitude) of the AC signal. It may take one or two periods to completely charge the capacitor. Once charged, the capacitor will not discharge. This is because any discharge would require a current to flow in the opposite direction which it cannot do because of the diode (Indeed current flows through the diode only during the *negative* cycle of the AC signal).

When completely charged, V_{out} will be a sum of the AC supply voltage at any particular moment *plus* the voltage across the capacitor. (Think of adding up potential differences around a Kirchhoff loop). An example input voltage and the resulting output voltage for this circuit is shown in Figures 1.5.6.1(h) and (i) respectively. As can be seen, the output has been shifted upwards by an amount equivalent to the capacitor voltage. (Note: the peak-to-peak voltage $V_{p\text{-}p}$ is twice the magnitude of V_p).

If the diode's direction is reversed, clearly the capacitor will charge up with an opposite polarity, and therefore the output V_{out} will be equivalent to the AC supply voltage at any particular moment *minus* the voltage across the capacitor.

Figures 1.5.6.2 shows how we can 'clamp' the AC signal to a chosen position by inclusion of a battery to the circuit. This introduces a greater flexibility into the design and allows clamping of the high and low peaks in the output to a particular level, without affecting the overall peak-to-peak value.

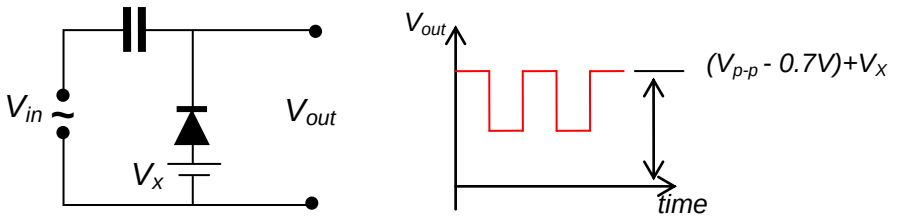

Figure 1.5.6.2: Diode clamp circuit with its output 'clamped' to a position dependent upon V_x. The figure shows (a) the circuit and (b) the output signal. The input signal is identical to that given in Figure 1.5.6.1 (h)

1.6 The transistor as a circuit element

We have invested quite a lot of effort in understanding how the simple p-n junction works in terms of charge mobility. This has helped us to appreciate how diodes function at a microscopic level and as a consequence we have been able to study in some detail various common applications of diodes, such as in power supplies and limiter circuits. Here we build on what we have learned over these introductory sections by turning our attention to a device with a far greater measure of electronic control, namely the *transistor*, which allows proportional control of a large current by a smaller, control current. In other words, the transistor enables *amplification*, which is essential for electronic circuits of all kinds (both analogue and digital) as we will find out throughout the remainder of these course notes.

Transistors are manufactured in two basic types: bipolar junction transistors (BJT) and field-effect transistors (FET). In the following sections we will first look at the properties of the BJT before moving on to examine the FET.

1.6.1 The bipolar junction transistor (BJT)

The bipolar junction transistor (BJT) is a device that employs both *p-type* and *n-type* semiconductor regions. Like the diode, these can be produced from a single crystal of silicon. There are two types of BJT, namely *npn transistor* and the *pnp transistor*. The npn transistor is basically a piece of p-material sandwiched between two pieces

of n-material and the pnp transistor is a piece of n-material sandwiched between two pieces of p-material. Here, we will consider only the npn type of transistor. (The pnp transistor has similar properties except that the current direction through the transistor is reversed). The npn bipolar junction transistor circuit symbol is shown in Figure 1.6.1.1, where the arrow in the diagram represents the direction of conventional current. (The circuit symbol for the pnp transistor is identical to the npn transistor symbol except the arrow is reversed). The circuit symbol has been labelled with the terms ‘collector’, ‘base’ and ‘emitter’ in order to make clear how it corresponds to the subsequent figures and description that follow.

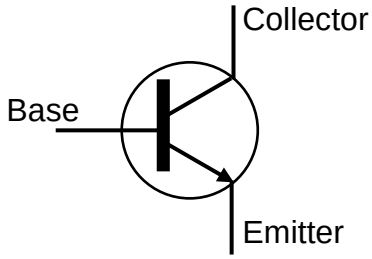

Figure 1.6.1.1: The circuit symbol of the npn bipolar junction transistor. The arrow indicates the direction of conventional current

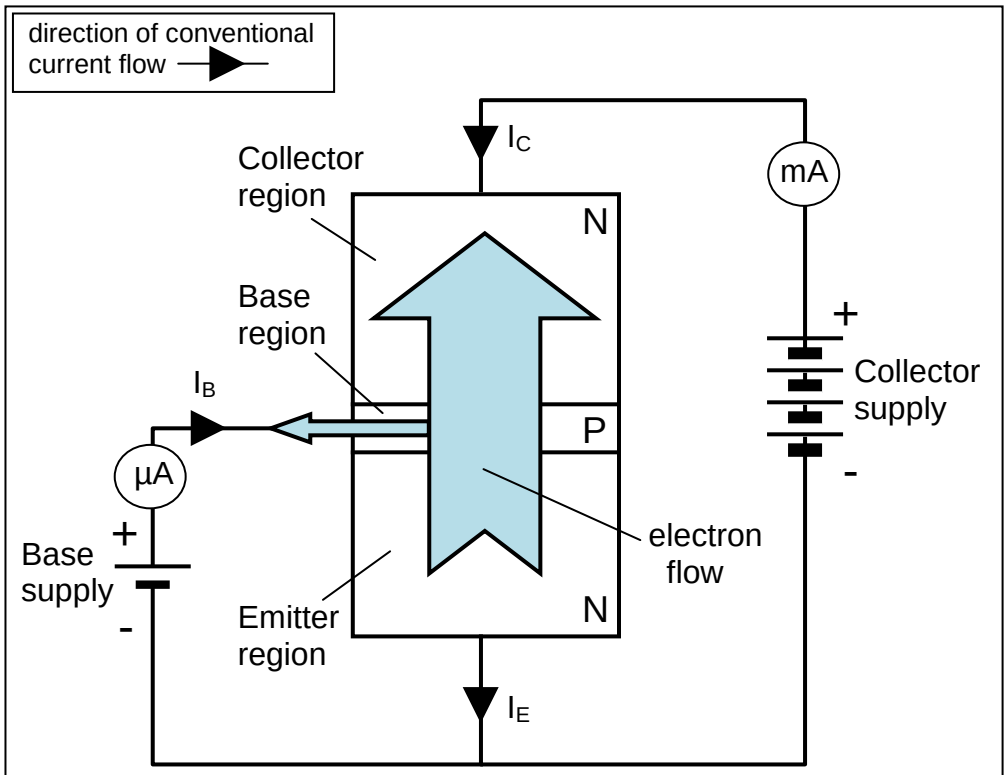

Figure 1.6.1.2: The npn bipolar junction transistor indicating directions of both conventional current flow and electron flow. Note that the base-emitter junction is forward-biased whilst the base-collector junction is reverse-biased

A detailed schematic of the npn transistor is shown in Figure 1.6.1.2. As can be seen, the 'n' and 'p' regions are labelled, along with the direction of conventional current flow and the direction of electron flow. This configuration is a typical 'test' configuration for measuring I_C as a function of I_B (known as the *forward transfer characteristic*). The region of p-material is called the *base*, whilst the two n-material regions are called the *collector* and *emitter* as indicated. The gain of the circuit is typically about 100, i.e. typically, $I_C = (100)I_B$.

With reference to Figure 1.6.1.2, in order for the transistor to function (either as an amplifier or indeed as a switch as we will see later), the base-emitter junction is forward-biased by the low voltage supply, whilst the base-collector junction is reverse-biased by the much larger supply. Electrons will therefore flow from the emitter to the base easily whilst there is very little flow of electrons to the base from the collector. In order to assist emitter-base electron flow, the emitter region is heavily doped and therefore has many electrons in its conduction band able to recombine with holes in the base region. Some electrons continue their journey out of the base to form the base current I_B. However, the base is actually very lightly doped with holes to *reduce* recombination. Moreover, it is made very narrow with a width typically about 1μm. This design causes the majority of electrons at the base-emitter junction to be attracted by the large positive potential connected to the collector. (If there are no available holes to recombine with in the base, the large collector potential will 'pull' the electrons up into the conduction band of the base, and thereon into the collector – refer to Figure 1.4.2(a) to help 'picture' this). Additionally, the collector region is also lightly doped which has the effect of increasing the width of the depletion region, and therefore effectively makes the p-type base even thinner.

In operation, around 99% of the electrons leaving the emitter (via emitter current I_E) arrive at the collector (hence a gain of about 100). Once the electrons are into the collector *n-type* material, they conduct as normal. Clearly if the base supply is turned off, then the collector current I_C would cease. In fact the collector current is proportional to the base current I_B, and hence its application as an amplifier. Moreover, the relationship between I_C and I_B is approximately linear within a designated *linear active region* which is the normal operational region for a particular transistor.

1.6.2 BJT amplifiers

A simple BJT voltage amplifier could be made by connecting a resistor R_C in series with the collector as shown in Figure 1.6.2.1. However, given that increased collector current will mean a larger voltage drop across R_C, and a smaller collector current a smaller voltage drop across R_C (from V=IR), the amplifier output will invert rises and falls in input voltage because $V_{out} = (V_+) - I_C R_C$. This is not suitable for all applications. Moreover, given what we have learned about the BJT so far, AC amplification will clearly not work with this particular circuit, since the current cannot reverse.

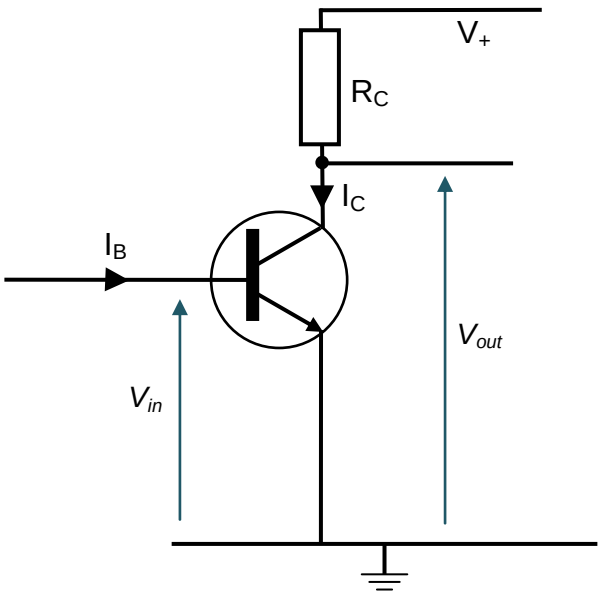

Figure 1.6.2.1: A resistor can be connected in series with the collector of the BJT to produce a voltage amplifier

However, we can adapt the circuit shown in Figure 1.6.2.1 to accommodate AC amplification by 'biasing' the circuit. Figure 1.6.2.2 shows such a circuit. The biasing of the circuit basically means that a DC voltage is *added* to the AC input voltage so that as the AC swings from high to low, it never actually falls to zero or a negative value, and therefore the current always flows in a single direction. (The AC voltage is oscillating above and below a positive voltage rather than the zero point).

Note: i_B, i_C and v_C all equal zero unless the AC voltage *Vin* is connected

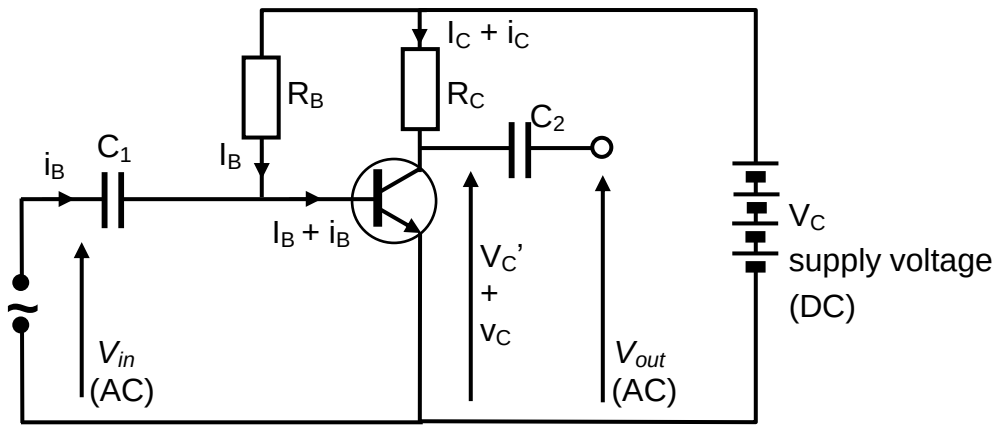

Figure 1.6.2.2: A 'biased' BJT circuit suitable for AC amplification. The particular amplifier circuit is known as a 'common-emitter amplifier'

With reference to Figure 1.6.2.2, the transistor bias current I_B is provided via the bias resistor R_B, connected between the base terminal and the collector supply voltage V_C. *With the amplifier operating under no input signal conditions* (and therefore i_B, i_C and v_C all equal zero), the value of R_B should be such that the resulting DC current I_C flowing in the collector resistor R_C produces a voltage drop across R_C equal to approximately 0.5 V_C Volts (half of the supply voltage).

The output voltage at the collector $V_C' = V_C - I_C R_C$ Volts. These values of collector current (I_C) and voltage (V_C') are known as the *quiescent* values (given that no AC signal is connected yet). With the output voltage at the collector set to half of the supply voltage, the AC output signal will be able to rise and fall equally on both negative and positive swings as mentioned above.

The function of capacitors C_1 and C_2 is to separate or block the DC bias voltages from the input signal source and any subsequent stages or loads connected to the output respectively. AC signals are 'coupled' through the capacitors but DC signals are not. (Read about 'coupling' for more information on this). In effect, the capacitors are blocking the DC because of their high reactance at low frequency – the basis of filter circuits that you have studied in a previous part of this module.

If the AC input signal is now connected to the amplifier via capacitor C_1, on the positive half cycle of the input signal, the forward bias of the base-emitter junction is increased, producing the additional AC base current i_B which now becomes in total (I_B+i_B), producing a subsequent increase (i_C) in collector current which now becomes (I_C+i_C).

Since the voltage at the collector $V_C' = V_C - (I_C+i_C)R_C$ volts at any time, the output voltage at the collector will *fall*.

However, on the negative half cycle of the input signal, the base current to the transistor will fall to a value (I_B-i_B). The collector current will consequently fall to a value (I_C-i_C), resulting in a *rise* in voltage at the collector.

The continuing change in collector voltage $v_C = i_C R_C$ volts, can be many times greater than the change in the input voltage – as explained above, about 100 times greater.

The BJT amplifier example shown in Figure 1.6.2.2 is known as the *common-emitter amplifier*. A different version of the amplifier, which we will not study here, is called the common-collector amplifier. The common-collector amplifier has its output connected to the emitter side of the transistor, across a resistor.

It should be noted that the AC output voltage is 180^o out of phase with the input signal since when the input voltage *rises*, the output *reduces* and *vice versa*, therefore the common-emitter amplifier is an *inverting amplifier*.

The waveforms of the common-emitter amplifier described above are shown in Figure 1.6.2.3.

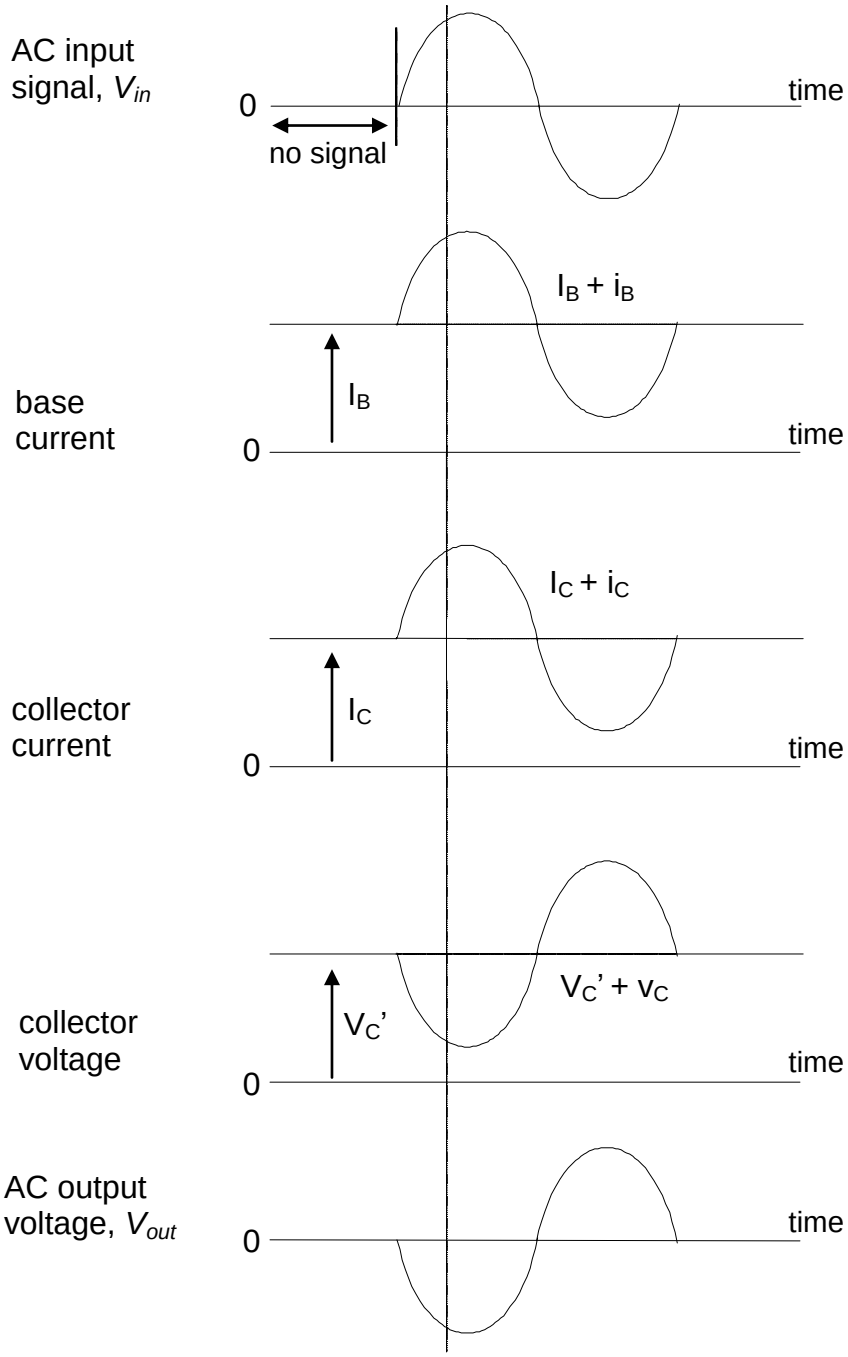

Figure 1.6.2.3: Waveforms of the BJT common-emitter amplifier circuit shown in Figure 1.6.2.2

1.6.3 Field-effect transistors – an overview

The field-effect transistor (FET) is similar to the BJT in some ways and dissimilar in others. For example, in both cases current can be controlled with another signal. But, in the case of the BJT, the control signal is a current, whilst in the FET the control signal is a voltage. Indeed the control input of an FET (called the 'gate') generally has a much higher input resistance than the BJT's base which makes a FET ideal for the input stage of a circuit. The FET input resistance is typically of the order MΩ or greater.

FETs are generally much less sensitive to temperature variations than BJTs, and consequently often more suitable for large-scale integrated circuits. However, FETs tend to have less gain than BJTs.

The FET has three connections, just like the BJT. The FET connections are the *drain*, *gate* and *source*, corresponding to the collector, base and emitter respectively of the BJT. Moreover, BJT circuits can constitute what are known as common-emitter amplifiers or common-collector amplifiers. Similarly, FET circuits can constitute a common-source amplifier or common-drain amplifier.

BJTs come in two types, pnp and npn. FETs are characterised by n-channel and p-channel types. However, FETs may have one of two types of gate construction: junction or metal oxide (JFET or MOSFET respectively). The metal oxide variety can also be doped in a particular way to become either depletion-mode or enhancement-mode.

A summary of the types of field-effect transistor discussed above is shown in Table 1.6.3.

Types of metal-oxide semiconductor field-effect transistor (MOSFET)	*Types of junction field-effect transistor (JFET)*
(i) MOSFET n-channel enhancement-mode	(i) JFET n-channel
(ii) MOSFET n-channel depletion-mode	(ii) JFET p-channel
(iii) MOSFET p-channel enhancement-mode	
(iv) MOSFET p-channel depletion-mode	

Table 1.6.3: Types of field-effect transistor

The analysis of all of the FETs described in Table 1.6.3 is very similar so it is not necessary to study every one individually. We will in fact consider the structure of three types of FET in the following sections (the n-channel types only) which will be sufficient to gain an overall understanding of all of the varieties. (A study of the p-channel type would be identical except that 'n' would become 'p', and voltages and currents would change polarity and direction respectively).

1.6.4 The junction field-effect transistor (JFET)

A schematic of the n-channel JFET is shown in Figure 1.6.4.1, along with its biasing voltages. The *gate* of the structure is basically a layer of p-type semiconductor material embedded on either side of a larger piece of n-type semiconductor. As mentioned above, the two ends of the n-type semiconductor are called the *source* and *drain* as can be seen. The drain/source voltage V_{DS} causes current to flow through the n-type material just as if it were a resistor, (i.e. the V-I characteristic is linear), unless either a gate/source voltage V_{GS} is applied or if V_{DS} becomes too large.

As can be seen, the negative gate voltage, which is connected to both p-type regions, causes a reverse bias of the p-n junctions, which in turn pass negligible current as a consequence (hence the high input resistance mentioned in the last section). The reverse bias of the p-n junctions causes there to be a depletion region at each, which grow in size as the reverse bias increases. The growth of depletion region effectively decreases the cross-sectional area of the n-type channel between the p-type regions, thus increasing its resistance to current flow.

Given that the potential along the length of the n-type channel has a gradient which varies from the 'source potential' at the source end, to V_{DS} volts at the drain, so too will the reverse bias vary along the length of the p-n junctions, and therefore also the size of the depletion region along the length of the junctions. On increasing the reverse bias, the depletion regions will eventually join together between the two p-type regions into a single depletion region, (first of all towards the *top* of the p-type regions in the figure), where reverse bias is greatest. This is called the *pinch-off point*. If V_{DS} is increased after the pinch-off point has been reached, the drain current

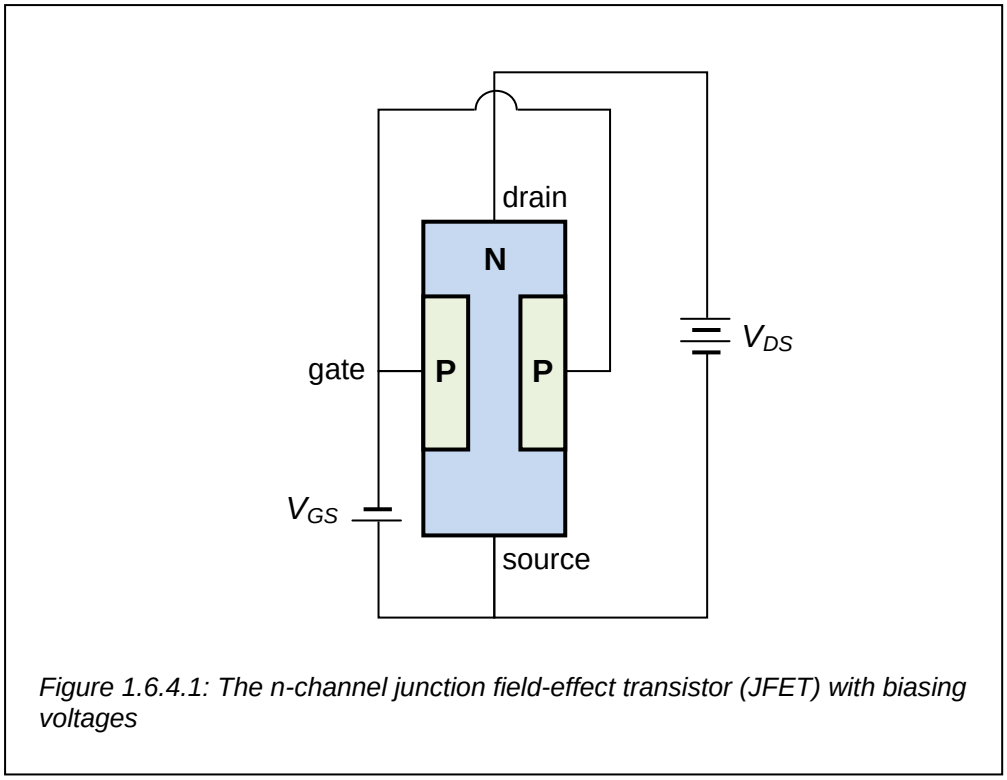

Figure 1.6.4.1: The n-channel junction field-effect transistor (JFET) with biasing voltages

remains approximately constant to reflect the balance between the increased V_{DS} and the consequential lower conductivity. However, normal operation would be at a reverse bias smaller than that which causes pinch-off. In this (linear) region a small change in gate voltage varies the drain current proportionally (a more negative voltage at the gate will increase reverse bias and therefore decrease the drain current). This is the property of the JFET that is used in the design of amplifiers. JFETs have many applications, for example the amplification of potentials produced by devices such as microphones. Moreover, given the nature of their high input resistance, they can be used to measure voltages for example in a digital test meter, since they draw negligible current and therefore do not affect the voltage that they are measuring.

The circuit symbols for both the n-channel and p-channel JFETs are shown in Figure 1.6.4.2.

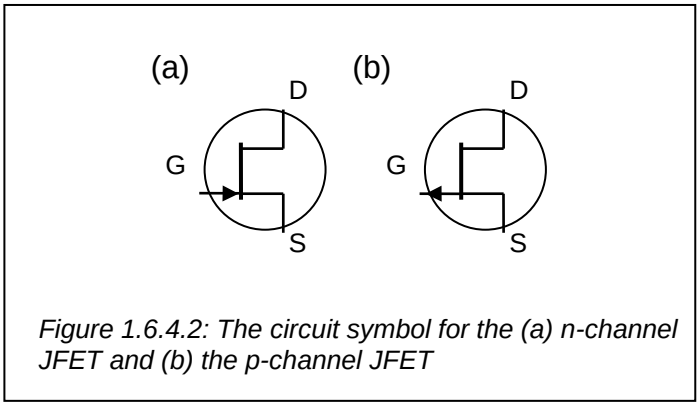

Figure 1.6.4.2: The circuit symbol for the (a) n-channel JFET and (b) the p-channel JFET

1.6.5 The metal oxide semiconductor field-effect transistor (MOSFET)

The JFET described in the previous section has recently been overtaken by the MOSFET as the most commonly-used field-effect transistor. The n-channel versions of the enhancement-mode and depletion-mode MOSFETs are shown in Figures 1.6.5.1 and 1.6.5.2 with typical biasing voltages. In the case of these transistors, the gate consists of a metal with good conductivity such as Aluminium, which is deposited upon an insulating material such as silicon dioxide. The insulator causes the DC input resistance to be extremely high.

In Figure 1.6.5.1, *the n-channel enhancement-mode MOSFET* has its source and drain connected to two pieces of n-type semiconductor, which are in turn embedded within a p-type substrate. Without a gate voltage, no current can flow between source and drain because the source-substrate junction and the substrate-drain junction are just like two diodes with opposite polarity, thus one is always reverse biased. However, when a *positive* voltage is applied to the gate, electron minority charge carriers (see section 1.3 above) from the p-type substrate migrate towards the gate creating a layer of charge called an *inversion layer*. The inversion layer of electrons enables current to flow between drain and source. If the gate voltage is increased, the amount of electrons in the inversion layer will increase, and consequently drain current will increase.

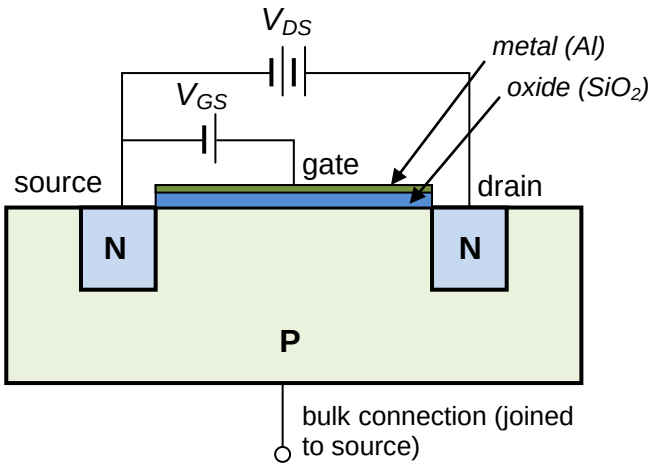

Figure 1.6.5.1: The n-channel enhancement-mode MOSFET with biasing voltages

The *n-channel depletion-mode MOSFET* shown in Figure 1.6.5.2, has some important differences to the enhancement-mode MOSFET. First of all, it already has an n-type channel below the gate along which current can flow normally as long as no gate voltage is applied. However, when a *negative* voltage is applied to the gate, the electrons in the n-type channel are repelled, decreasing available electrons for conduction and consequently the drain current decreases. Indeed the versatility of this particular transistor is further increased by the fact that if a *positive* voltage is applied to the gate, even more electrons will be attracted to the channel region, increasing current above the level seen with no voltage applied at all.

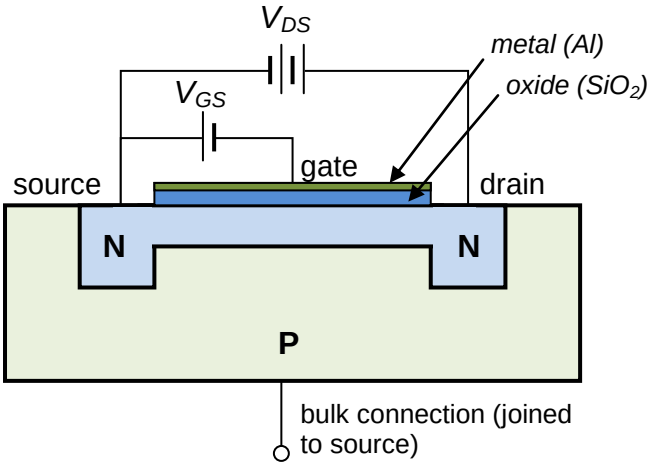

Figure 1.6.5.2: The n-channel depletion-mode MOSFET with biasing voltages

Circuit symbols for the various MOSFET types are shown in Figure 1.6.5.3.

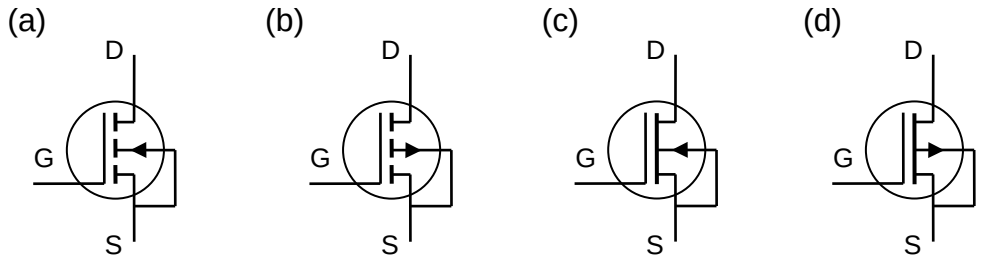

Figure 1.6.5.3: The circuit symbol for the (a) n-channel enhancement-mode MOSFET, (b) p-channel enhancement-mode MOSFET (c) n-channel depletion-mode MOSFET and (d) p-channel depletion-mode MOSFET

1.6.6 The MOSFET common-source amplifier

A typical transistor amplifier circuit, known as the *MOSFET common-source amplifier* is shown in Figure 1.6.6. Typical values of circuit components have been included. With a fixed supply voltage, the magnitude of the drain current depends only upon the voltage difference between the gate and the source, V_{GS}. Below a certain threshold voltage of V_{GS} which is normally about one or two volts, current does not flow. This is the voltage required to form the inversion layer of electrons mentioned in the previous section. (There is a potential gradient along the prospective channel between source and drain within the substrate so the channel of electrons will not necessarily be formed along the entire length at voltages below threshold. This 'potential gradient' effect was mentioned in Section 1.6.4 in the case of the JFET).

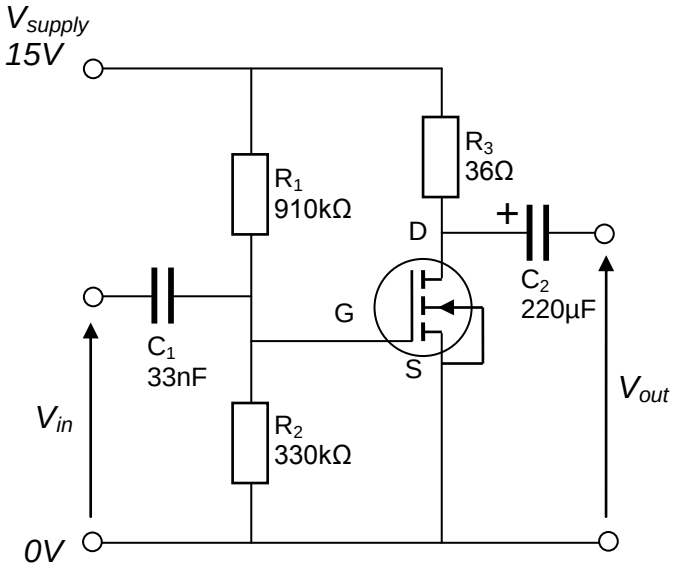

Figure 1.6.6: The n-channel enhancement-mode MOSFET common-source amplifier

As can be seen in the circuit of Figure 1.6.6, the gate of the MOSFET is held at a quiescent (fixed) voltage, as we saw in the case of the BJT amplifier. The quiescent voltage is determined by the resistors R_1 and R_2. C_1 couples the amplifier to a previous circuit, for example a microphone. The amplifier circuit then converts the input voltage into an output (drain) current. The low value of R_3 ensures that the drain current is maximised, which could be useful for driving e.g. a speaker. C_2 couples the output signal to the load whilst not letting the DC quiescent current pass. If the concept of 'coupling' is not familiar, a good way of understanding how the capacitor blocks DC is to realise that when a load is connected to V_{out}, C_2 and the load act together to form a high-pass filter.

One final thing to note is that as in the case of the BJT common-emitter amplifier shown previously, when used to amplify an AC voltage, the MOSFET common-source amplifier shown here, is an *inverting amplifier* since the greater the drain current, the greater the voltage drop across R_3, and consequently the smaller the output voltage V_{out}.

NOTE: The transistors introduced and described in these notes can be used as 'building blocks' for much more complicated circuits. The introduction of more than one transistor into the circuit can increase current and voltage gain for applications where high current or voltage is needed. For example, the BJT 'Darlington Pair' consists of two BJTs wired together and is packaged into a single chip. The current gain of the pair is equal to the product of the current gains of the individual transistors.

Where high power amplification is important, for example to drive an audio amplifier to produce sound at high volume, or to drive motors which actuate industrial robots, there are many configurations possible which are beyond the scope of these course notes. Power amplifiers include multiple transistors amongst other components, and have specialised designs to minimise noise, overheating etc.

1.6.7 Transistor switches

In addition to their use as amplifiers, transistors are used extensively as *switches*, e.g. to turn on and off lamps, sirens and motors. The transistor becomes useful when the control current is small (i.e. of the order μA) and the current needed for the lamp, siren, motor etc. is much larger. This makes it possible to feasibly control high current devices from sensors, logic gates (see later in these notes) or other low-current circuits.

Figure 1.6.7.1 gives an example of a typical MOSFET switching circuit which causes a lamp to be automatically illuminated when a *photodiode* detects low levels of light.

A photodiode is a reverse-biased diode which is manufactured such that it has an optical window through which light can pass to become incident on the depletion region of the diode. A 'particle' of light known as a photon will liberate a valence electron, and therefore also a hole, when the electron absorbs the photon's energy. If the energy absorbed is greater than the forbidden band gap between valence and conduction bands, the electron will 'jump' to the conduction band and will drift away from the junction thus creating a 'photocurrent'. This type of current is similar to the 'reverse leakage current' discussed earlier in the notes where electrons were thermally excited before being swept along by the pd across the depletion region, although the photocurrent has a much greater magnitude in comparison. The current produced by incident photons is proportional to the overall light incident on the photodiode hence the photodiode's ability to measure light intensity.

With reference to Figure 1.6.7.1, when a typical photodiode is illuminated by say, average room lighting, the photodiode's photocurrent will rise to a few milliamps (mA) or more. The photocurrent will pass through the resistor R causing a voltage drop across R. The value of R can therefore be chosen such that the voltage across R (from V=IR) results in the potential between R and the photodiode being below the threshold for MOSFET operation (see in previous section), and therefore the transistor will be 'off' and the lamp not lit.

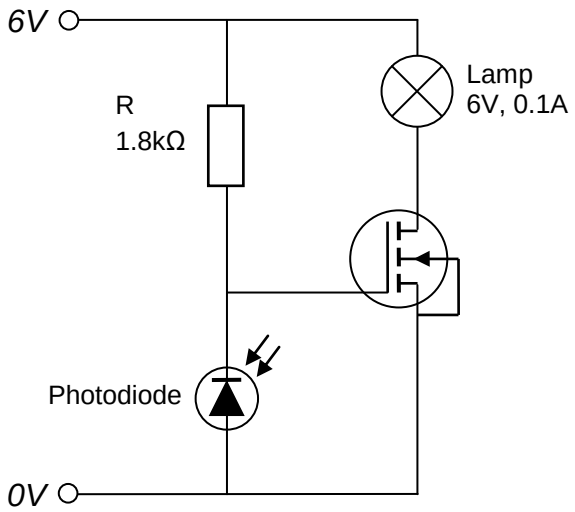

Figure 1.6.7.1: A MOSFET switching circuit. The lamp illuminates when the light level falls below a given level

However, as light levels reduce, the photocurrent reduces and therefore also the voltage across R. This causes the potential between R and the photodiode to rise beyond the MOSFET threshold input voltage (a couple of volts) and the lamp will be lit. The transistor will in fact be driven into *saturation*. This is typical in switching circuits in order for them to work most effectively. (The saturation point is beyond the

'linear' regime used in amplification circuits. It is the point where the drain current cannot be increased further due to the doping structure of the transistor, even though the input voltage may increase). If there is no light incident on the photodiode the current reverts back to the 'reverse leakage-current' level which is of the order microamps (μA). Indeed as described in an earlier section, the MOSFET can be turned on with virtually no current present - it is the input *voltage* that is important. The MOSFET is therefore ideal for a circuit controlled by a photodiode – the BJT for example, may not have received sufficient current to function correctly.

Another switching circuit which includes a BJT is shown in Figure 1.6.7.2. The circuit acts as a switch for a mains supply heater, and could be used for e.g. maintaining a particular temperature in a room. There are two components in the circuit other than the supply, resistors, a diode and the BJT: a relay for switching on and off the mains voltage and an NTC thermistor which acts as the sensor in the circuit.

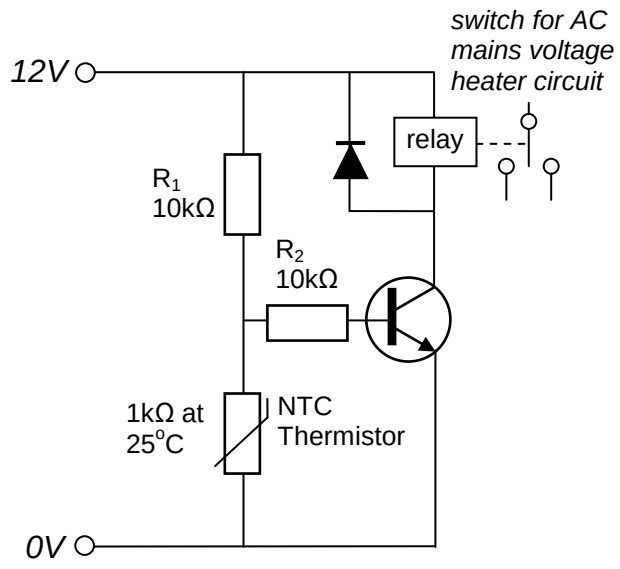

Figure 1.6.7.2: A BJT switching circuit for a mains supply heater

A relay is basically a coil of wire wound on an iron core. As we have seen in an earlier part of the course, (AC circuit analysis, and with reference to an inductor), a coil of wire will produce a magnetic field when current is passed through it. The magnetic field that is produced in the relay's coil is enhanced by the presence of the iron core. When the relay is energised, the magnetic field will attract a contact arm known as an armature (the particular mechanism depending on the design of the relay) which causes contacts to be closed thus completing the circuit of the higher voltage supply and load. The relay switch is controlled by a small voltage but is able to turn on and off much larger supplies. For example, relay circuits are used throughout a car, for the horn, lights, windscreen-wipers etc., and for mains supply appliances such as air-conditioning.

A negative temperature coefficient (NTC) thermistor is basically a temperature-sensitive resistor. In fact all resistors do vary slightly with temperature but NTC thermistors are particularly sensitive as they are made of semiconductor material. We know that thermal energy can allow a current to flow more readily through a semiconductor material due to its small band-gap. This means that resistance decreases in semiconductors with increasing temperature and this is the case for NTC thermistors also i.e. the resistance of an NTC thermistor decreases with increasing temperature. A typical coefficient of resistance with temperature for an NTC thermistor is -4%/^{o}C, and has a typical operational range of -50 to 300^{o}C.

There are in fact positive temperature coefficient (PTC) thermistors, which are generally designed to have a very steep increase in resistance at a particular temperature and are often used for protection circuitry to sense for example, too much current.

With reference to Figure 1.6.7.2, if room temperature is high, then the resistance of the NTC thermistor is low, and therefore the voltage between R1 and the thermistor is also low, and as a consequence both the base current and the collector current through the BJT are also low – in fact too low to produce a sufficient reaction in the relay to operate the heater switch. However, if the room temperature drops, the thermistor's resistance will increase and therefore the voltage across the thermistor will increase also. This in turn will cause the base and collector currents to increase, eventually beyond the threshold necessary to move the relay's armature, which will cause the heater's switch to be closed and the heater to be turned on.

Finally, again with reference to Figure 1.6.7.2, there is one component that doesn't appear to have played a role so far: the diode wired in parallel to the relay. In fact this diode is wired into the circuit in order to protect the transistor, and is known as a *protective diode*.

When the transistor turns off, the current will cease to flow through the relay which will induce an emf across the relay as the magnetic field present within the coil of the relay collapses.

The emf thus induced can be hundreds of volts. This level of voltage can produce a current of sufficient magnitude to permanently damage a transistor. The diode wired in parallel to the relay provides a path for the current produced by the relay's coil at switch-off so that it does not pass through the transistor. Note that the orientation of the diode is such that no current will pass through it during the transistor's normal operation.

If you are not familiar with how inductors affect current when switches are turned on and off read about 'DC transients'. Basically, a coil or inductor will produce an emf across itself which will try to oppose changes in the current flowing through it. When the current is turned on, a 'back emf' is induced across the coil which prevents the current from reaching its expected magnitude immediately. The current will rise gradually. When current is turned off, an emf is once again induced across the coil which causes current to continue flowing for some time whilst gradually reducing in magnitude to zero.