

Analogue Electronics 7: Non-linear components – Diodes and Transistors

Here I am going to introduce *two new components* and give a short outline of the semiconductor concepts behind them. Both are **non-linear components**, i.e. the relationship between I and V that they introduce *cannot*, in general, be described as a change in *amplitude* and *phase*. If you imagine an electric current as a flowing fluid then the **diode** is equivalent to a *valve* and the **transistor** is equivalent to a *tap*. The diode still is a **passive component**, while the transistor is an **active component**, it amplifies power.

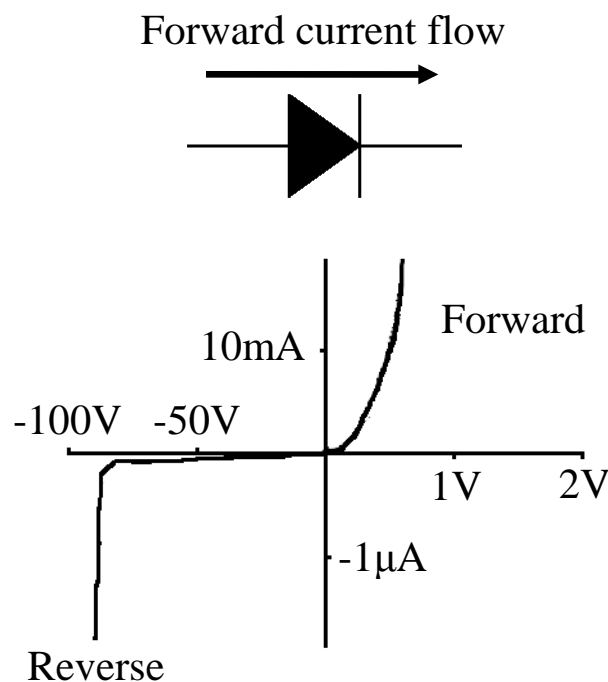
Diodes as non-linear components: (H&H, 1.25, p. 44)

The **diode** is a *passive, non-linear* component. To first order it is a one-way conductor (with an accompanying voltage drop). Being a non-linear component has the following consequences:

- a diode doesn't obey Ohm's law,
- circuits with diodes have no Thevenin equivalent.

You may well have been using light emitting diodes (LEDs) with your flip-flops in checkpoint D2. The LEDs light up when a current flows.

The symbol for a diode and the relationship between current and voltage, the **characteristic IV-curve**, are shown below:

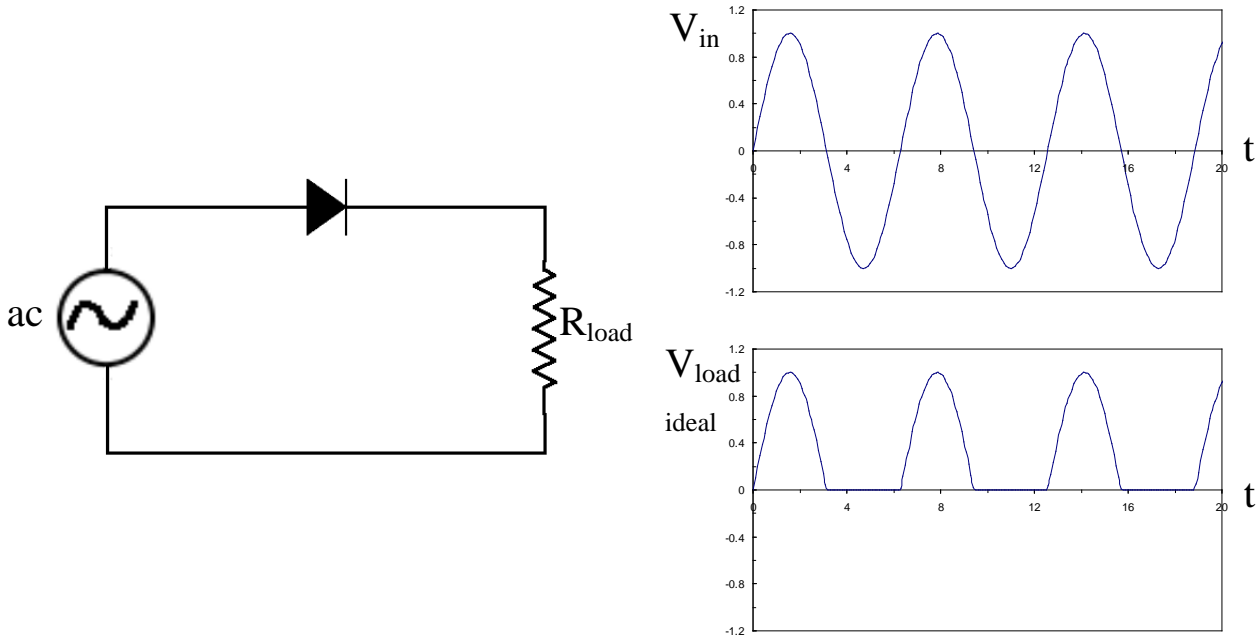


Take note of the characteristic scales involved for the behaviour in **forward** and **backward direction**: the curve above illustrates a factor of 50 between the positive and negative voltage scale and a change in four orders of magnitude in the current scale.

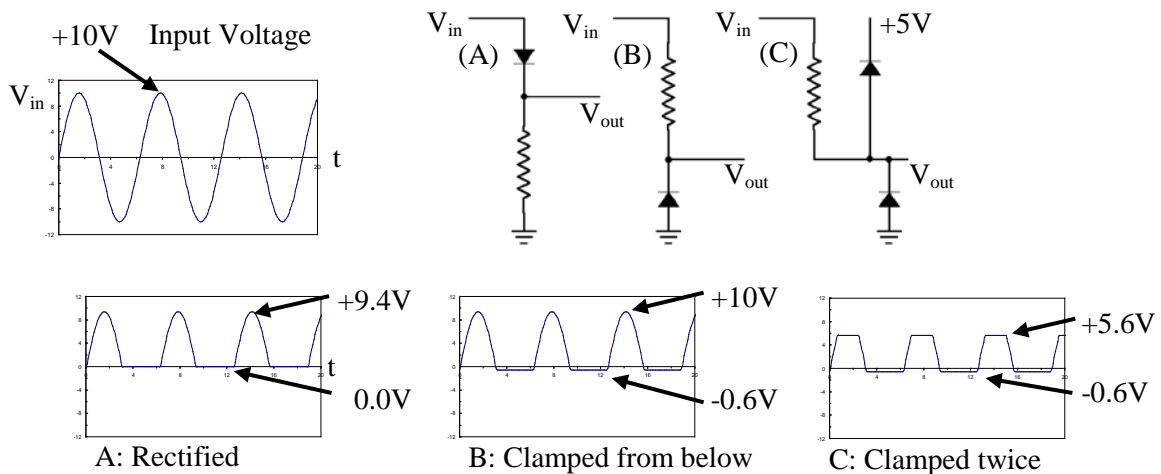
A diode is not a perfect one way conductor. It behaves like a valve only if the input voltage is substantially larger than the **forward voltage drop** of about $\sim 0.6V$: below that voltage the IV-curve features a characteristic shape, above the diode becomes conductive and the voltage drop across the diode is constant (up to the maximum current rating the diode can sustain). On the reverse side the (mirrored) shape of the curve actually is very similar, but on very different scales and the diode cannot sustain any significant current flow beyond the **break-through point** (defining the maximum reverse bias rating of the diode, about -80V in the example above) – it very quickly will be destroyed, i.e. it will blow up.

Diodes as rectifiers: (H&H, 1.26, p. 44)

In the circuit on the left below a diode is used to regulate the output of an oscillating function generator. The input voltage, above graph on the right, oscillates between positive and negative values. In an *ideal world* only the positive part of the input voltage would appear across the load, see the lower graph on the right:



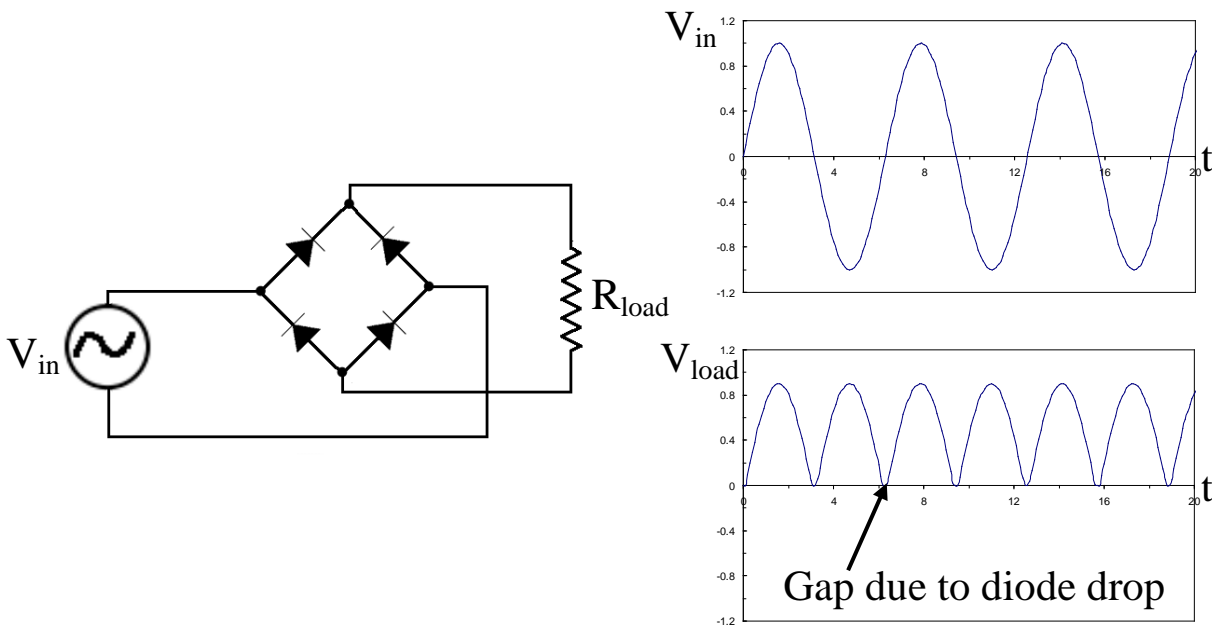
But it is also necessary to consider the characteristic “diode drop” of $\sim 0.6V$ which takes place when a diode conducts. Below the effect of three different voltage divider configurations are shown involving the use of one or more diodes. The input signal is a sine wave with amplitude of 10V. Roughly 0.6V get lost as voltage drop across the diode. Using this information and understanding the voltage divider you should be able to understand the three output graphs below. It is very clear that a diode doesn’t obey Ohm’s law.



The diode drop is the main feature which makes the diode different from an ideal valve for current. Diodes are used in many applications (if you do the “chaos” experiment in the second semester – you will put one in an LCR circuit). One of the common applications is in turning an oscillating supply voltage into a DC voltage.

Full-wave bridge rectifier: (H&H, 1.26 & 1.27, p. 44-46)

The circuit (A) pictured on the previous page can be considered a very simple **rectification circuit**. A voltage which oscillates between positive and negative is input and only the positive part is output. This is extremely inefficient because half of the input voltage is lost. With a more advanced design *both halves* of the input voltage can be recovered. In the circuit below diodes are arranged head to head and tail to tail on four sides of a square, with the pick-up on the other two corners that the feed-in voltage. Whatever the sign of the input voltage, a positive voltage appears across the load, as illustration below.



Two diodes are always *in series* with the input regardless of whether it is in a positive or negative oscillation. The graphs next to the circuit show the input wave form (top) and the output of the circuit (bottom). Both halves of the oscillations have been recovered. The only *loss* is due to the diode drop. Note that we have recovered a time dependent positive voltage. For most DC applications you would actually want a steady voltage rather than a **series of ripples**.

In a practical full wave rectifier there are *further components* between the diodes and the load. The first is a *capacitor*. This stores charge as the voltage increases and releases it in the gaps when the voltage drops. The result is to smooth out the big dips between voltage peaks. However, the voltage will still have a slight ripple. The final flattening out is carried out using a **Zener diode**. In contrast to standard diodes they allow for a controlled current flow at their breakdown voltage. Thus, a wide range of current can flow at an *almost fixed voltage* and hence one can get rid of the final ripple. If you are interested in this you can read further about Zener diodes in the library or on the internet.

Semiconductors: (Brophy, Basic Electronics for Scientists, ch. 4, p. 97)

Around the middle of the twentieth century first circuits were made using non-linear components. Components behaving like valves then were made out of *vacuum tubes* (similar to old fashion TV tubes). Some of the first computers, e.g. ENIAC, used such vacuum tubes (but not the very first computer, Z3, which used only linear relays...). For a computing power which today appears completely pathetic their size was that of a house and the amount of heat generated was excessive – let alone the perpetual task of swapping burnt vacuum tubes. As quantum mechanics became better understood it was realized that a *semiconductor* was a good analogue of a vacuum tube. This led to the semiconductor revolution, widely available computers and hence the possibility of generating quantitative understanding on a previously un-envisaged scale. Related Nobel Prizes are:

- 1956 – Shockley, Bardeen & Brattain “for their researches on semiconductors and their discovery of the transistor effect”
- 2000 – Alferov, Kroemer & Kilby “for basic work on information and communication technology”

Today exciting developments with semiconductors involve *plastic electronics*, *quantum dots*, *carbon nanotubes* and *spintronics*. I am going to leave all of that and just stick with simple ideas about *inorganic crystalline solids*.

The behaviour of electrons in a crystalline solid is described by the **Fermi model**. Each energy state in the crystal can only be filled by a *single electron* (this is due to the fermionic nature of the electrons). Due to the huge number of electrons in a macroscopic crystal the structure of energy levels becomes a *near continuum*. But there can be energy gaps between distinct **energy bands**. At the absolute zero temperature, $T=0\text{K}$, *all states* are occupied from the lowest up to a maximum, called the **Fermi level**, above which all other states remain free. The highest energy band which contains occupied states is called **valence band**. The lowest band containing free states is called **conductive band**.

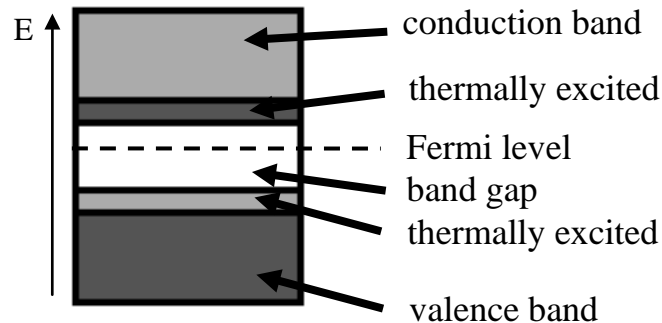
In **metals** valence and conductive band are the same as the Fermi level lies in the *middle of an energy band*. With increasing temperature some states get *thermally excited*, generating a discrete number of free states below and filled states above the Fermi level. Electrons which are excited above the Fermi level can move through the crystal **quasi-free** in the conductive band. At the same time the vacated states, often referred to as **holes**, can move in the valence band. The more electrons are excited the higher the *conductivity* of the material becomes. At room temperature in crystalline metals there already is an abundance of excited electrons, which is why their resistance is very low. (But the resistance rises again with the temperature as now the scattering with the thermally moving ions in the lattice have overtaken the scarcity of charge carriers as the dominant source for the resistance.)

By contrast, in an **insulator** the Fermi level resides in a *big energy gap* between the valence band and the conductive band. At $T=0\text{K}$ the valence band is *completely filled* and the conductive band is *completely empty*. The probability for thermal excitement of electrons across the band gap decreases exponentially with its size. At room temperature the conductivity of insulators can be *24 orders of magnitude smaller* than that of metals, and is relatively independent of the temperature.

In **semiconductors** the Fermi level also resides in a band gap, but here the *band gap is small*. A sizable number of electrons can jump from the valence to the conducting band by thermal excitation, enough to provide *some conductivity*, even in a pure semiconductor crystal. This

conductivity is significantly distinct by orders of magnitude to either side from that of metals and insulators, respectively. In this regime the scarcity of charge carriers still limits the conductivity of the material, which is why the resistance decreases with rising temperature. A simplified illustration of the energy band situation for a semiconductor is shown below.

Simple energy-band model for a semiconductor



In this cartoon the energy of electron states increases as you proceed vertically up the diagram. The darker gray regions are electron states that are filled. Some electrons have been thermally excited over the band gap into the conduction band. Hence there are empty states (holes) in the valence band, and filled states in the conduction band. There are no states in the band gap.

In practical devices the presence of mobile charge carriers is not left down to thermal excitation alone. *Higher densities* of charge carriers, up to several orders of magnitude, can be created by **doping** the semiconductor with **impurities** of a different valence (called **donors** when they provide additional *electrons* and **acceptors** if they provide additional *holes*). The impurity atoms lead to a decrease in the size of the band gap and a permanent population of mobile charge carriers. Depending on the valence of the dopant there will either be an excess or a shortfall of electrons. If there is an *excess of electrons* then the material will preferentially conduct via the **motion of electrons** and is called an **n-type** semiconductor (n for negative). If there is a *shortage of electrons* then the material will conduct via the **motion of holes** through the valence band and is called a **p-type** semiconductor (p for positive). The carriers that are available in excess are called the **majority charge carriers**. For a given sample of semiconductor the sign of the majority carriers can be determined via their behaviour in a perpendicular magnetic field, making use of the *Hall effect*. In summary:

n-type:	majority electrons	dopant: provides donor atoms	excess of electrons
p-type:	majority holes	dopant: provides acceptor atoms	shortage of electrons

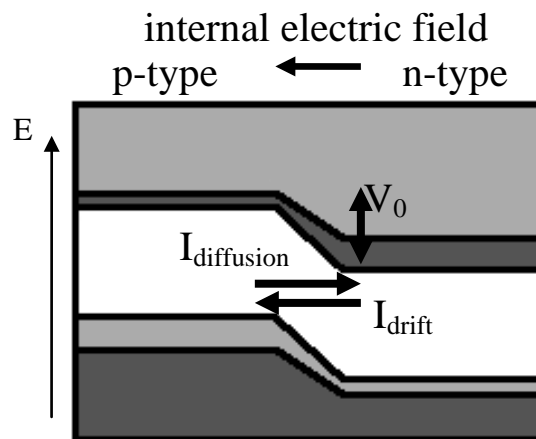
For the semiconductor revolution to take place there was one more ingredient needed.

pn-junction: (Brophy, Basic Electronics for Scientists, ch. 4, p. 101)

In order to build devices with new functionality pieces of semiconductor with different properties, i.e. different doping, need to be combined. Ideally this is done in a single crystal semiconductor, but the techniques to achieve that are beyond the scope of this lecture. All fundamental characteristics can be discussed by the means of the **pn-junction**, the most basic and most common type of artificial semiconductor structure.

In the **pn-junction** p-type and n-type layers of material are prepared *right next to each other* in the semiconductor crystal. So a region with an excess of electrons meets a region with a shortage of electrons (which in the first place both are electrically neutral). Naturally the

electrons start to diffuse to equalise the electron concentration by filling the holes across the junction (that increases the entropy). This diffusion of charge locally creates *electrically charged up regions* on either side of the junction which also turn *depleted of free carriers of charge*. As the **depletion region** grows by the diffusion of charge carriers so does the *electrical field* across the depletion region, until this field stops further charge from crossing the pn-junction. (This is why the *depletion region* also often is called the *space charge region*.) A **dynamical equilibrium** is formed between the force which causes the *electron diffusion* due to the different electron concentrations and the counter-acting force from the *electrostatic field* caused by the charge separation. At the same time electrons drift from n-type to p-type material *holes drift in the opposite direction* just adding to the described effects. Also: once the *majority charge carriers* from either side cross the pn-junction they become called the **minority charge carriers** on the other side. The resulting equilibrium state of a pn-junction is illustrated in the energy-band model below.



There is a *potential difference* V_0 between the energy levels of the valence and conductor band of the p-type and n-type material. This corresponds to the strength of the internal electric field caused by the diffusion of charge. Electrons diffuse from n-type to p-type material, holes (and hence the conventional current) diffuse in the opposite direction. Together they form the **diffusion current**, $I_{diffusion}$. The charge carriers drifting due to the internal electric field form the **drift current**, I_{drift} . Left to form an equilibrium both currents *cancel*. But the balance between $I_{diffusion}$ and I_{drift} can be *disturbed by applying an external voltage* across the junction.

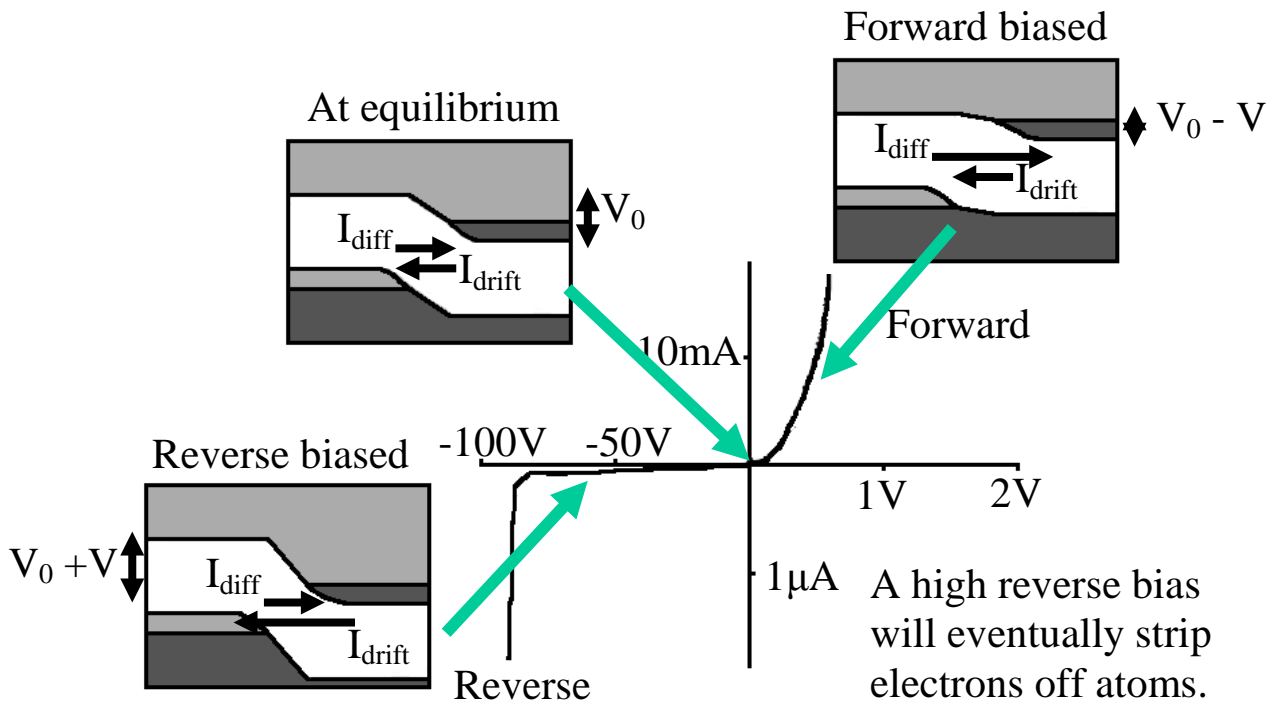
The semiconductor diode:

You may have seen this coming: **the pn-junction actually is the diode**. By applying an external voltage across the junction one finds the behaviour discussed above for a diode. This is illustrated further in the graph on the next page.

When the external voltage is applied in **forward direction**: as the external voltage decreases the internal electrical field and hence the drift current is reduced. Up to an applied voltage of the value of the forward voltage drop *the difference* between the diffusion and the drift current appears as *external current* through the diode. At the forward voltage drop the external voltage *compensates fully* the internal electrical field. Beyond that point *additional current* can flow through the junction as defined by the external circuitry up to some structural limit, e.g. by thermal stresses.

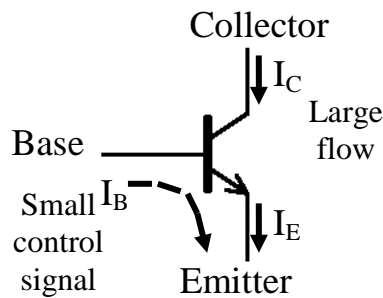
When the external voltage is applied in **reverse direction**: as the external voltage increases the internal electrical field and hence the drift current increases. *The difference* between the diffusion and the drift current appears as *external current* through the diode, but this is *extremely small*. Effective charge transport through the junction is blocked by the increasing

internal electrical field, i.e. by the increased potential difference between the two materials. The picture only changes at the *break-through point* again. There the internal field becomes strong enough to *rip further electrons off the semiconductor lattice*. A massive conductance sets in, which very quickly reaches the structural limits of the material, leading to a catastrophic failure of the device.



Introduction to transistors: (H&H, 2.01 & 2.02, p. 61-65)

One can regard a **transistor** to be equivalent to a tap: one can *control* the amount of current which flowing through it. A **Bipolar Junction Transistor (BJT)** consists of two *pn-junctions* back-to-back, either in a npn- or in a pnp-structure. The graph below shows the symbol of a BJT npn-transistor together with labels and indications of the current flow.

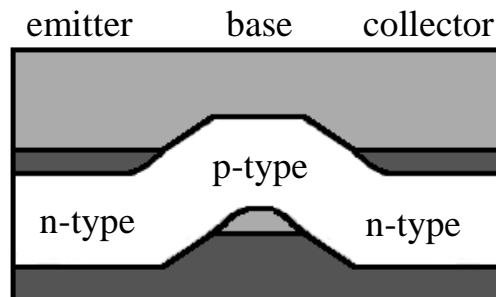


A *small current* is directed into the device via its **Base** connection (the central p-type layer in the example). With the Collector being connected to an external current supply this gives rise to a *large current* flowing from the **Collector** to the **Emitter** (here both of n-type). Thus, the transistor *amplifies power*. The gain lies between the small control current and the large Collector-To-Emitter current. The gain in signal is what characterises an **active component**.

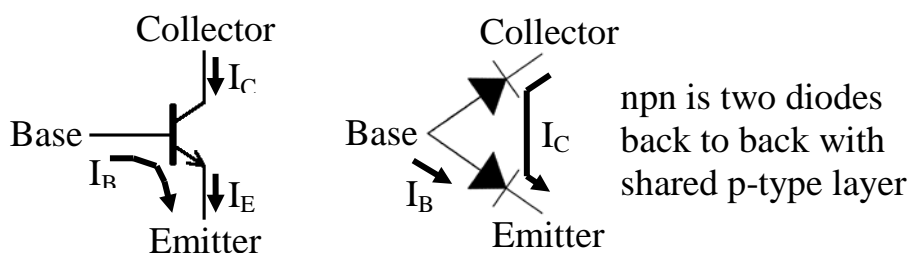
Transistors are the principal component of integrated circuits, and hence they are of terrific scientific and cultural importance. By dealing with the constituents of integrated circuits we are working our way back towards the logic chips that we began with. Remember: the protocol for logic “0”s and “1”s we are using in the lab is called TTL (Transistor-Transistor-Logic) giving some indication that it is transistors that are inside all of those logic gates.

Bipolar Junction Transistor: (Brophy, Basic Electronics for Scientists, ch. 4, p. 108)

Actually, there are two main classes of transistors: the already mentioned **Bipolar Junction Transistors** (BJT) and the **Field Effect Transistors** (FET). This lecture has not the room to discuss the characteristics of the many variants of the BJT and FET, but will demonstrate the basic behaviour by the means of the *npn-variant of the BJT*. A typical energy-band structure of such a transistor, in its equilibrium, is shown below.



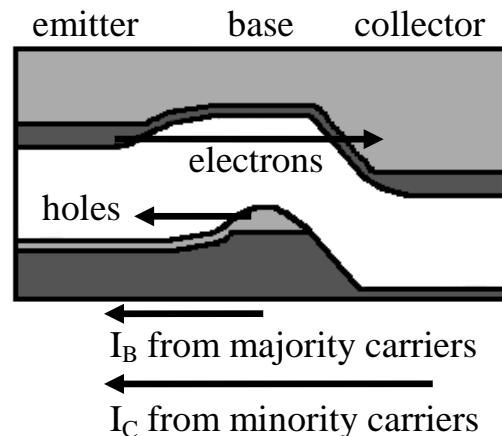
The key point to note is that the *two back-to-back pn-junctions share the p-type volume*, which actually is only a *thin layer*. This is like two diodes being build back-to-back on the same substrate. The following cartoon illustrates this picture.



Evidently, when using *two discrete diodes* one would *not* get the amplification behaviour of the transistor. The upper diode is reverse biased and blocks any significant current from collector to emitter. So what is different in the transistor giving us conduction in this arrangement? The *shared, thin p-type volume* is the key. In order to reach the active (i.e. amplification) region of the BJT *two conditions* have to be fulfilled:

- 1) The junction between B and E needs to be forward biased: $V_B - V_E \approx 0.6V$
- 2) V_C has to be larger than V_E by at least 0.2V

The corresponding energy-band structure looks like this.

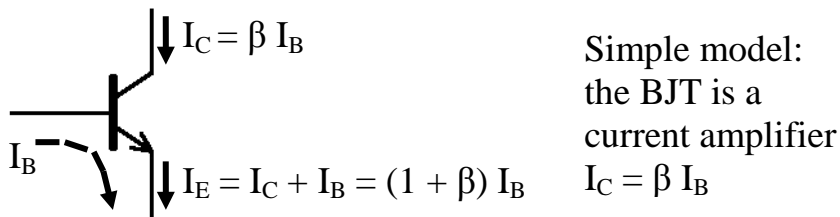


A small steering current injected into the base, I_B , will lead to the *majority carriers* of the base, the holes, flowing through the forward biased diode, towards the emitter. In return electrons

are injected from the emitter into the base region, where they become *minority carriers*. Since the base region is thin, only about one in hundred electrons will recombine with a hole in the base, which is forming the **base current**, I_B . The others diffuse further across the reverse biased junction between base and collector, effectively in the wrong direction. As the collector is positively biased with respect to the emitter the electrons are carried away, emerging as **collector current**, I_C . This current crosses directly from the one n-type region to the other. Because both the *positive majority* and *negative minority carriers* contribute to the **emitter current**, I_E , these transistors are called **bipolar** (in contrast FET operate *unipolar*). For the above conditions one finds

$$I_C = \beta I_B \quad \text{with } \beta \approx 100, \text{ while } I_E = I_C + I_B$$

where β is the current gain. This *simple model* is illustrated in the following picture.



In some cases one may even further simplify and *neglect the base current* as it is small with respect to the collector current. Then one gets:

- emitter current equals collector current: $I_E = I_C$
- voltage between base and emitter: $V_{BE} \approx 0.6V$

We will use that in the next example, a very basic device to amplify voltage.

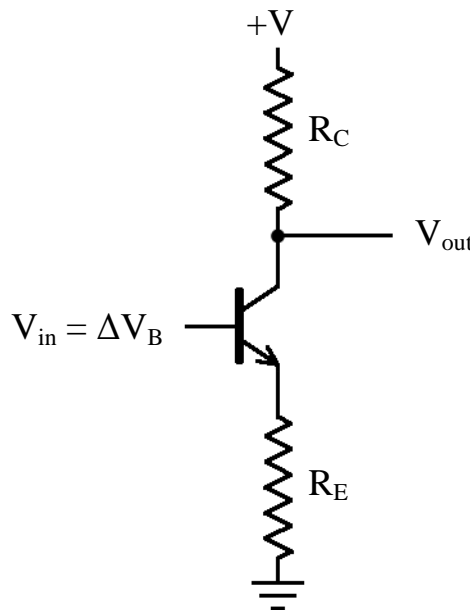
In general the BJT is a complicated beast with lots of frequency and temperature dependent non-ideal behaviour. The beast is tamed only by additional circuitry around it, controlling its behaviour for a given working range of parameters, usually at the expense of gain. Well balanced integrated circuits with guaranteed operational parameters then are referred to as **operational amplifiers**, or **OpAmp** for short. Though for integration reasons these days many of such circuits are realised using MOSFET transistors. This is what you will be working with in the lab.

We will look at a single example where the second – simpler – model is sufficient to understand the device. Our example will be a device for amplifying voltage. It is called a common-emitter amplifier.

BJT example: common-emitter amplifier (H&H, 2.07, p. 76)

The actual behaviour of a transistor *depends on the circuitry* around it, e.g. it can be configured to amplify voltage or current or both. In this example, the **common-emitter amplifier**, we look at a very basic way to *amplify voltage*. The simplest model of the transistor, the one above even *ignoring the base current*, is sufficient to get a proper description for it.

Consider the circuit fragment shown on the next page. The input voltage is fed to the base, the output voltage is measured at the collector. Both signals and the emitter share the common ground, hence the name of this circuit topology. So far we have looked at the transistor effect in terms of the flow of charges and currents. We need to establish what happens to voltages



for the configuration of resistors and connections shown.

A signal $V_{in} = \Delta V_B$ arrives at the base connection. We have to assume that with any variation the absolute potential is still held above the diode drop.

This results in a variation of the voltage to ground of ΔV_B at the emitter. The emitter current I_E flows with these variations:

$$\Delta I_E = \frac{\Delta V_E}{R_E} = \frac{\Delta V_B}{R_E}$$

and will be large if R_E is small. Further we approximate the collector current to: $I_C \approx I_E$. This current is flowing from the power supply through the collector and emitter and to ground. The voltage drops along the way will depend on the size of the current and the resistances. Due to R_C any changes in the current I_C result in a change in V_{out} . If R_C is large then ΔV_{out} is large:

$$\Delta V_{out} = -\Delta I_C R_C = -\Delta V_B \left(\frac{R_C}{R_E} \right)$$

The output voltage is an inverted, amplified version of the input voltage ΔV_B . The degree of amplification in this circuit to first order is only controlled by the *relative sizes of the external resistors* and it can provide a large gain. Still it has serious limitations: one is limited to small input signals to avoid signal distortions, further dependencies on temperature and bias current show outside narrow windows of operational parameters. But his is enough about the nuts and bolts of transistors. Next we will move on to higher level devices.