Chapter 29

Electronics - Grade 12

29.1 Introduction

Electronics and electrical devices surround us in daily life. From the street lights and water pumps to computers and digital phones, electronics have enabled the digital revolution to occur. All electronics are built on a backbone of simple circuits, and so an understanding of circuits is vital in understanding more complex devices.

This chapter will explain the basic physics principles of many of the components of electronic devices. We will begin with an explanation of capacitors and inductors. We see how these are used in tuning a radio. Next, we look at **active** components such as transistors and operational amplifiers. Lastly, the chapter will finish with an explanation of digital electronics, including logic gates and counting circuits.

Before studying this chapter, you will want to remind yourself of:

- The meaning of voltage (V), current (I) and resistance (R), as covered in Grade 10 (see chapter 10), and Grade 11 (see chapter 19).
- Capacitors in electric circuits, as covered in Grade 11 (see section 17.6).
- Semiconductors, as covered in Grade 11 (see chapter 20).
- The meaning of an alternating current (see section 28.3).
- Capacitance (C) and Inductance (L) (see section 28.4).

29.2 Capacitive and Inductive Circuits

Earlier in Grade 12, you were shown alternating currents (a.c.) and you saw that the voltage and the current varied with time. If the a.c. supply is connected to a resistor, then the current and voltage will be proportional to each other. This means that the current and voltage will 'peak' at the same time. We say that the current and voltage are **in phase**. This is shown in Figure 29.1.

When a capacitor is connected to an alternating voltage, the maximum voltage is proportional to the maximum current, but the maximum voltage does not occur at the same time as the maximum current. The current has its maximum (it peaks) one quarter of a cycle before the voltage peaks. Engineers say that the 'current leads the voltage by $90°$'. This is shown in Figure 29.2.

For a circuit with a capacitor, the instantaneous value of $\frac{V}{I}$ is not constant. However the value of $\frac{V_{\text{max}}}{I_{\text{max}}}$ is useful, and is called the **capacitive reactance** (X_C) of the component. Because it is still a voltage divided by a current (like resistance), its unit is the ohm. The value of X_C (C

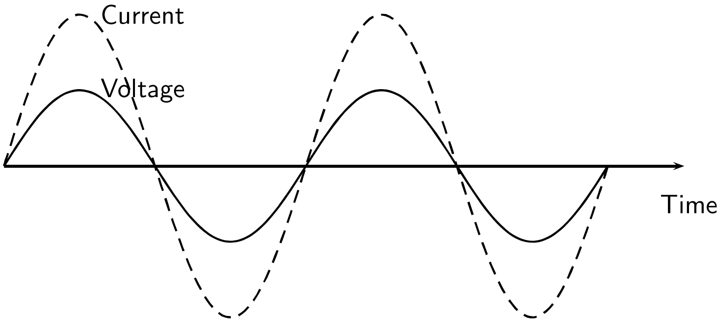


Figure 29.1: The voltage and current are in phase when a resistor is connected to an alternating voltage.

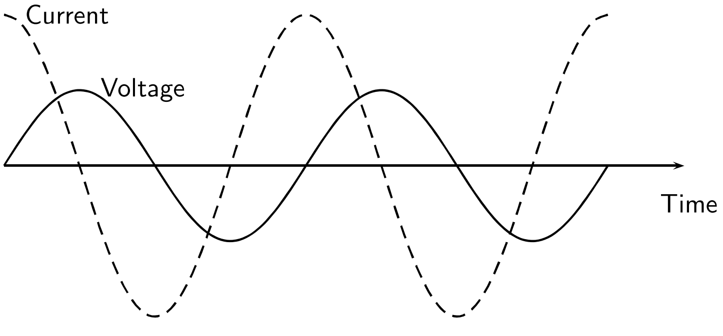


Figure 29.2: The current peaks (has its maximum) one quarter of a wave before the voltage when a capacitor is connected to an alternating voltage.

standing for capacitor) depends on its capacitance (C) and the frequency (f) of the alternating current (in South Africa 50 Hz).

$$X_C = \frac{V_{\text{max}}}{I_{\text{max}}} = \frac{1}{2\pi f C} \tag{29.1}$$

Inductors are very similar, but the current peaks $90°$ after the voltage. This is shown in Figure 29.3. Engineers say that the 'current lags the voltage'. Again, the ratio of maximum voltage to maximum current is called the reactance — this time inductive reactance (X_L). The value of the reactance depends on its inductance (L).

$$X_L = \frac{V_{\text{max}}}{I_{\text{max}}} = 2\pi f L \tag{29.2}$$

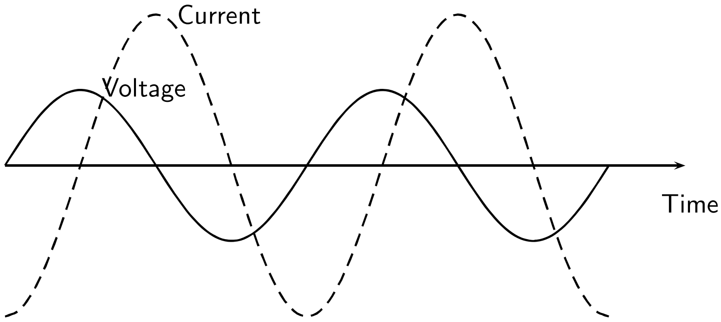


Figure 29.3: The current peaks (has its maximum) one quarter of a wave after the voltage when an inductor is connected to an alternating voltage.

Definition: Reactance
The ratio of the maximum voltage to the maximum current when a capacitor or inductor is connected to an alternating voltage. The unit of reactance is the ohm.

While inductive and capacitive reactances are similar, in one sense they are opposites. For an inductor, the current peaks $90°$ **after** the voltage. For a capacitor the current peaks $90°$ **ahead** of the voltage. When we work out the total reactance for an inductor and a capacitor in series, we use the formula

$$X_{\text{total}} = X_L - X_C \tag{29.3}$$

to take this into account. This formula can also be used when there is more than one inductor or more than one capacitor in the circuit. The total reactance is the sum of all of the inductive reactances minus the sum of all the capacitive reactances. The magnitude (number) in the final result gives the ratio of maximum voltage to maximum current in the circuit as a whole. The sign of the final result tells you its phase. If it is positive, the current peaks $90°$ after the voltage, if it is negative, the current peaks $90°$ before the voltage.

If a series circuit contains resistors as well, then the situation is more complicated. The maximum current is still proportional to the maximum voltage, but the phase difference between them won't be $90°$. The ratio between the maximum voltage and maximum current is called the **impedance** (Z), and its unit is also the ohm. Impedances are calculated using this formula:

$$Z = \sqrt{X^2 + R^2} \tag{29.4}$$

where X is the total reactance of the inductors and capacitors in the circuit, and R is the total resistance of the resistors in the circuit.

It is easier to understand this formula by thinking of a right angled triangle. Resistances are drawn horizontally, reactances are drawn vertically. The hypotenuse of the triangle gives the impedance. This is shown in Figure 29.4.

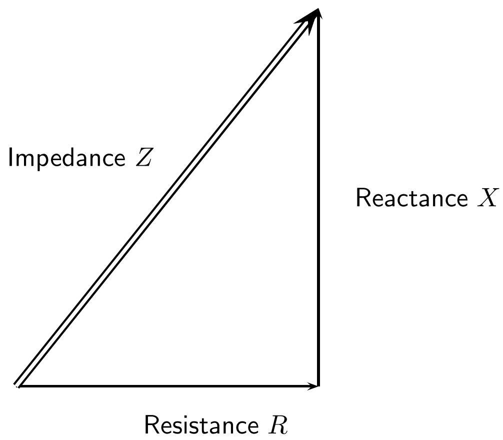


Figure 29.4: Visualizing the relationship between reactance, resistance and impedance.

Definition: Impedance
The maximum voltage divided by the maximum current for any circuit. The unit of impedance is the ohm.

It is important to remember that when resistors and inductances (or capacitors) are in a circuit, the current will not be in phase with the voltage, so the impedance is not a resistance. Similarly the current won't be exactly $90°$ out of phase with the voltage so the impedance isn't a reactance either.

Worked Example 179: The impedance of a coil

Question: Calculate the maximum current in a coil in a South African motor which has a resistance of 5 Ω and an inductance of 3 mH. The maximum voltage across the coil is 6 V. You can assume that the resistance and inductance are in series.

Answer

1. Calculate the reactance of the coil $X_L = 2\pi f L = 2\pi \times 50 \times 0{,}003 = 0{,}942\ \Omega$
2. Calculate the impedance of the coil
   $Z = \sqrt{X^2 + R^2} = \sqrt{0{,}942^2 + 5^2} = 5{,}09\ \Omega$

3. Calculate the maximum current $I_{\max} = V_{\max}/Z = 6/5{,}09 = 1{,}18$ A.

Worked Example 180: An RC circuit

Question: Part of a radio contains a 30 Ω resistor in series with a 3 μF capacitor. What is its impedance at a frequency of 1 kHz?

Answer

1. Calculate the reactance of the capacitor

$$X_C = \frac{1}{2\pi f C} = \frac{1}{2\pi \times 10^3 \times 3 \times 10^{-6}} = 53{,}05\ \Omega$$

2. Calculate the impedance $Z = \sqrt{X^2 + R^2} = \sqrt{53{,}05^2 + 30^2} = 60{,}9\ \Omega$

Exercise: Capacitive and Inductive Circuits

1. Why is the instantaneous value of $\frac{V}{I}$ of little use in an a.c. circuit containing an inductor or capacitor?
2. How is the reactance of an inductor different to the reactance of a capacitor?
3. Why can the ratio of the maximum voltage to the maximum current in a circuit with a resistor and an inductor not be called a reactance?
4. An engineer can describe a motor as equivalent to a 30 Ω resistor in series with a 30 mH inductor. If the maximum value of the supply voltage is 350 V, what is the maximum current? Assume that the frequency is 50 Hz.
5. A timer circuit in a factory contains a 200 μF capacitor in series with a 10 kΩ resistor. What is its impedance? Assume that the frequency is 50 Hz.
6. A 3 mH inductor is connected in series with a 100 μF capacitor. The reactance of the combination is zero. What is the frequency of the alternating current?

Interesting Fact

Most factories containing heavy duty electrical equipment (e.g. large motors) have to pay extra money to their electricity supply company because the inductance of the motor coils causes the current and voltage to get out of phase. As this makes the electricity distribution network less efficient, a financial penalty is incurred. The factory engineer can prevent this by connecting capacitors into the circuit to reduce the reactance to zero, as in the last question above. The current and voltage are then in phase again. We can't calculate the capacitance needed in this chapter, because the capacitors are usually connected in parallel, and we have only covered the reactances and impedances of series circuits.

29.3 Filters and Signal Tuning

29.3.1 Capacitors and Inductors as Filters

We have already seen how capacitors and inductors pass current more easily at certain frequencies than others. To recap: if we examine the equation for the reactance of a capacitor, we see that the frequency is in the denominator. Therefore, when the frequency is low, the capacitive reactance is very high. This is why a capacitor blocks the flow of DC and low frequency AC because its reactance increases with decreasing frequency.

When the frequency is high, the capacitive reactance is low. This is why a capacitor allows the flow of high frequency AC because its reactance decreases with increasing frequency.

Therefore putting a capacitor in a circuit blocks the low frequencies and allows the high frequencies to pass. This is called a high pass filter. A filter like this can be used in the 'treble' setting of a sound mixer or music player which controls the amount of high frequency signal reaching the speaker. The more high frequency signal there is, the 'tinnier' the sound. A simple high pass filter circuit is shown in Figure 29.5.

Similarly, if we examine the equation for the reactance of an inductor, we see that inductive reactance increases with increasing frequency. Therefore, when the frequency is low, the inductive reactance is very low. This is why an inductor allows the flow of DC and low frequency AC because its reactance decreases with decreasing frequency.

When the frequency is high, the inductive reactance is high. This is why an inductor blocks the flow of high frequency AC because its reactance increases with increasing frequency.

Therefore putting an inductor in a circuit blocks the high frequencies and allows the low frequencies to pass. This is called a low pass filter. A filter like this can be used in the 'bass' setting of a sound mixer or music player which controls the amount of low frequency signal reaching the speaker. The more low frequency signal there is, the more the sound 'booms'. A simple low pass filter circuit is shown in Figure 29.6.

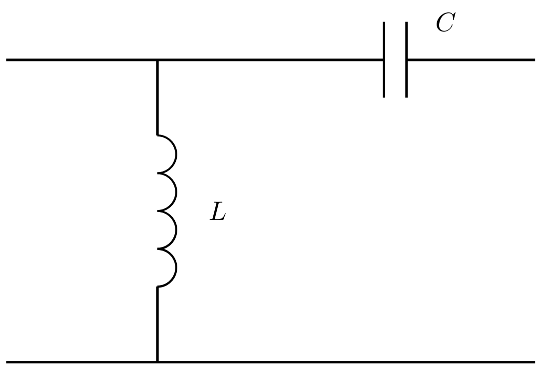


Figure 29.5: A high pass filter. High frequencies easily pass through the capacitor and into the next part of the circuit, while low frequencies pass through the inductor straight to ground.

29.3.2 LRC Circuits, Resonance and Signal Tuning

A circuit containing a resistor, a capacitor and an inductor all in series is called an LRC circuit. Because the components are in series, the current through each component at a particular time will be the same as the current through the others. The voltage across the resistor will be in phase with the current. The voltage across the inductor will be $90°$ ahead of the current (the current always follows or lags the voltage in an inductor). The voltage across the capacitor will be $90°$ behind the current (the current leads the voltage for a capacitor). The phases of the three voltages are shown in Figure 29.7.

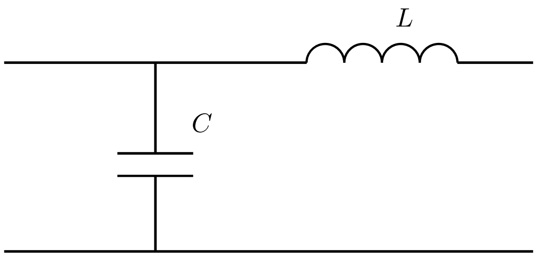


Figure 29.6: A low pass filter. Low frequencies pass through the inductor and into the next part of the circuit, while high frequencies pass through the capacitor straight to ground.

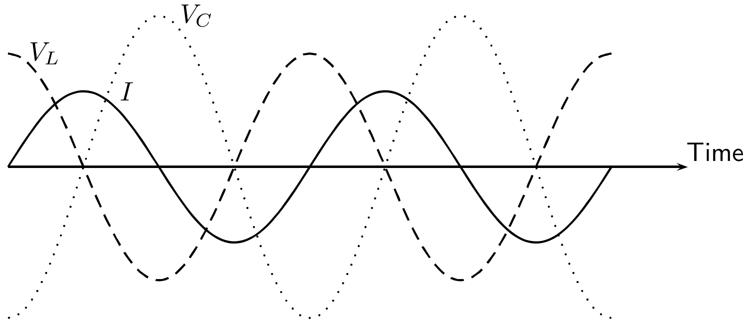


Figure 29.7: The voltages across the separate components of an LRC circuit. Looking at the peaks, you see that the voltage across the inductor V_L 'peaks' first, followed $90°$ later by the current I, followed $90°$ later by the voltage across the capacitor V_C. The voltage across the resistor is not shown — it is in phase with the current and peaks at the same time as the current.

The reactance of the inductor is $2\pi fL$, and the reactance of the capacitor is $1/2\pi fC$ but with the opposite phase. So the total reactance of the LRC circuit is

$$X = X_L - X_C = 2\pi fL - \frac{1}{2\pi fC}$$

The impedance of the circuit as a whole is given by

$$Z = \sqrt{X^2 + R^2} = \sqrt{\left(2\pi fL - \frac{1}{2\pi fC}\right)^2 + R^2}$$

At different frequencies, the impedance will take different values. The impedance will have its smallest value when the positive inductive reactance cancels out the negative capacitive reactance. This occurs when

$$2\pi fL = \frac{1}{2\pi fC}$$

so the frequency of minimum impedance must be

$$f = \frac{1}{2\pi\sqrt{LC}}$$

This is called the **resonant frequency** of the circuit. This is the frequency at which you can get the largest current for a particular supply voltage. It is also called the **natural frequency** of the circuit. This means the frequency at which the circuit would oscillate by itself.

Definition: Resonance
Resonance occurs when a circuit is connected to an alternating voltage at its natural frequency. A very large current in the circuit can build up, even with minimal power input.

An LRC circuit is very useful when we have a signal containing many different frequencies, and we only want one of them. If a signal like this is connected to an LRC circuit, then only the resonant frequency (and other frequencies close to it) will drive measureable currents. This means that an LRC circuit can select one frequency from a range. This process is called **signal tuning**.

Interesting Fact

When you set up a radio antenna, and amplify the radio signal it receives, you find many different bands of frequencies — one from each radio station. When you listen to the radio, you only want to listen to one station. An LRC circuit in the radio (the tuning circuit) is set so that its resonant frequency is the frequency of the station you want to listen to. This means that of the many radio stations broadcasting, you only hear one. When you adjust the tuning dial on the radio, it changes the capacitance of the capacitor in the tuning circuit. This changes the resonant frequency, and this changes the radio station that you will hear.

Exercise: Filters and Signal Tuning

1. Which component would you use if you wanted to block low frequencies?

2. Which component would you use if you wanted to block high frequencies?
3. Calculate the impedance of a series circuit containing a 50 Ω resistor, a 30 μF capacitor and a 3 mH inductor for frequencies of (a) 50 Hz, (b) 500 Hz, and (c) 5 000 Hz.
4. Calculate the resonant frequency of the circuit in the last question.
5. A radio station broadcasts at a frequency of 150 kHz. The tuning circuit in the radio contains a 0.3 mH inductor. What is the capacitance of the capacitor needed in the tuning circuit if you want to listen to this radio station?
6. State the relationship between the phase of the voltages across an inductor, a resistor and a capacitor in an LRC circuit.
7. Explain what is meant by resonance.
8. Explain how LRC circuits are used for signal tuning, for example in a radio.

29.4 Active Circuit Elements

The components you have been learning about so far — resistors, capacitors and inductors — are called **passive** components. They do not change their behaviour or physics in response to changes in voltage or current. **Active** components are quite different. Their response to changes in input allows them to make amplifiers, calculators and computers.

29.4.1 The Diode

A diode is an electronic device that allows current to flow in one direction only.

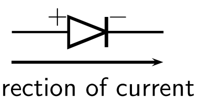


Figure 29.8: Diode circuit symbol and direction of flow of current.

A diode consists of two doped semi-conductors joined together so that the resistance is low when connected one way and very high the other way.

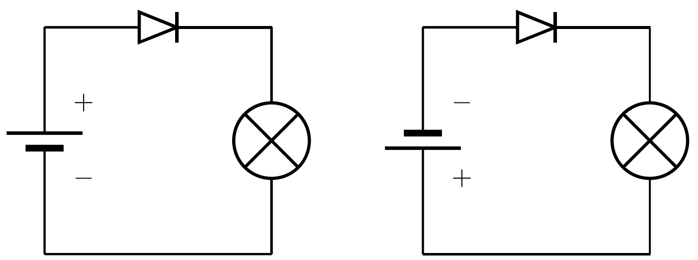


Figure 29.9: Operation of a diode. (a) The diode is forward biased and current is permitted. The negative terminal of the battery is connected to the negative terminal of the diode. (b) The diode is reverse biased and current flow is not allowed. The negative terminal of the battery is connected to the positive terminal of the diode.

A full explanation of diode operation is complex. Here is a simplified description. The diode consists of two semiconductor blocks attached together. Neither block is made of pure silicon — they are both **doped**. Doping was described in more detail in Section 10.3.

In short, p-type semiconductor has fewer free electrons than normal semiconductor. 'P' stands for 'positive', meaning a lack of electrons, although the material is actually neutral. The locations where electrons are missing are called **holes**. This material can conduct electricity well, because electrons can move into the holes, making a new hole somewhere else in the material. Any extra electrons introduced into this region by the circuit will fill some or all of the holes.

In n-type semiconductor, the situation is reversed. The material has an more free electrons than normal semiconductor. 'N' stands for 'negative', meaning an excess of electrons, although the material is actually neutral.

When a p-type semiconductor is attached to an n-type semiconductor, some of the free electrons in the n-type move across to the p-type semiconductor. They fill the available holes near the junction. This means that the region of the n-type semiconductor nearest the junction has no free electrons (they've moved across to fill the holes). This makes this n-type semiconductor positively charged. It used to be electrically neutral, but has since lost electrons.

The region of p-type semiconductor nearest the junction now has no holes (they've been filled in by the migrating electrons). This makes the p-type semiconductor negatively charged. It used to be electrically neutral, but has since gained electrons.

Without free electrons or holes, the central region can not conduct electricity well. It has a high resistance, and is called the **depletion band**. This is shown in Figure 29.10.

You can explain the high resistance in a different way. A free electron in the n-type semiconductor will be repelled from the p-type semiconductor because of its negative charge. The electron will not go into the depletion band, and certainly won't cross the band to the p-type semiconductor. You may ask, "But won't a free electron in the p-type semiconductor be attracted across the band, carrying a current?" But there are no free electrons in p-type semiconductor, so no current of this kind can flow.

If the diode is reverse-biased, the + terminal of the battery is connected to the n-type semiconductor. This makes it even more negatively charged. It also removes even more of the free electrons near the depletion band. At the same time, the − terminal of the battery is connected to the p-type silicon. This will supply free electrons and fill in more of the holes next to the depletion band. Both processes cause the depletion band to get wider. The resistance of the diode (which was already high) increases. This is why a reverse-biased diode does not conduct.

Another explanation for the increased resistance is that the battery has made the p-type semiconductor *more negative* than it used to be, making it repel any electrons from the n-type semiconductor which attempt to cross the depletion band.

On the other hand, if the diode is forward biased, the depletion band is made narrower. The negative charge on the p-type silicon is cancelled out by the battery. The greater the voltage used, the narrower the depletion band becomes. Eventually, when the voltage is about 0,6 V (for silicon) the depletion band disappears. Once this has occurred, the diode conducts very well.

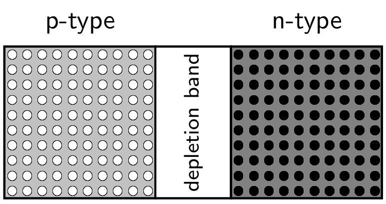


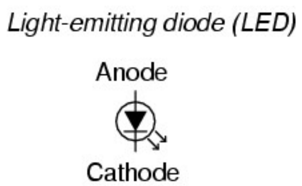
Figure 29.10: A diode consists of two doped semi-conductors joined together so that the resistance is low when connected one way and very high the other way.

Exercise: The Diode

1. What is a diode?
2. What is a diode made of?
3. What is the term which means that a diode is connected the 'wrong way' and little current is flowing?
4. Why is a diode able to conduct electricity in one direction much more easily than the other?

29.4.2 The Light Emitting Diode (LED)

A light-emitting diode (LED) is a diode device that emits light when charge flows in the correct direction through it. If you apply a voltage to force current to flow in the direction the LED allows it will light up.



Extension: Circuit Symbols

This notation of having two small arrows pointing away from the device is common to the schematic symbols of all light-emitting semiconductor devices. Conversely, if a device is light-activated (meaning that incoming light stimulates it), then the symbol will have two small arrows pointing toward it. It is interesting to note, though, that LEDs are capable of acting as light-sensing devices: they will generate a small voltage when exposed to light, much like a solar cell on a small scale. This property can be gainfully applied in a variety of light-sensing circuits.

The color depends on the semiconducting material used to construct the LED, and can be in the near-ultraviolet, visible or infrared part of the electromagnetic spectrum.

Interesting Fact

Nick Holonyak Jr. (1928) of the University of Illinois at Urbana-Champaign developed the first practical visible-spectrum LED in 1962.

Light emission

The wavelength of the light emitted, and therefore its color, depends on the materials forming the p-n junction. A normal diode, typically made of silicon or germanium, emits invisible far-infrared light (so it can't be seen), but the materials used for an LED have emit light corresponding to near-infrared, visible or near-ultraviolet frequencies.

LED applications

LEDs have many uses. Some of these are given here.

- thin, lightweight message displays, e.g. in public information signs (at airports and railway stations, among other places)
- status indicators, e.g. on/off lights on professional instruments and consumers audio/video equipment
- infrared LEDs in remote controls (for TVs, VCRs, etc)
- clusters of LEDs are used in traffic signals, replacing ordinary bulbs behind colored glass
- car indicator lights and bicycle lighting
- calculator and measurement instrument displays (seven segment displays), although now mostly replaced by LCDs
- red or yellow LEDs are used in indicator and [alpha]numeric displays in environments where night vision must be retained: aircraft cockpits, submarine and ship bridges, astronomy observatories, and in the field, e.g. night time animal watching and military field use
- red or yellow LEDs are also used in photographic darkrooms, for providing lighting which does not lead to unwanted exposure of the film
- illumination, e.g. flashlights (a.k.a. torches, UK), and backlighting for LCD screens
- signaling/emergency beacons and strobes
- movement sensors, e.g. in mechanical and optical computer mice and trackballs
- in LED printers, e.g. high-end color printers

LEDs offer benefits in terms of maintenance and safety.

- The typical working lifetime of a device, including the bulb, is ten years, which is much longer than the lifetimes of most other light sources.
- LEDs fail by dimming over time, rather than the abrupt burn-out of incandescent bulbs.
- LEDs give off less heat than incandescent light bulbs and are less fragile than fluorescent lamps.
- Since an individual device is smaller than a centimetre in length, LED-based light sources used for illumination and outdoor signals are built using clusters of tens of devices.

Because they are monochromatic, LED lights have great power advantages over white lights where a specific color is required. Unlike the white lights, the LED does not need a filter that absorbs most of the emitted white light. Colored fluorescent lights are made, but they are not widely available. LED lights are inherently colored, and are available in a wide range of colors. One of the most recently introduced colors is the emerald green (bluish green, about 500 nm) that meets the legal requirements for traffic signals and navigation lights.

The largest LED display in the world is 36 m high, at Times Square, New York, U.S.A.

There are applications that specifically require light that does not contain any blue component. Examples are photographic darkroom safe lights, illumination in laboratories where certain photo-sensitive chemicals are used, and situations where dark adaptation (night vision) must be preserved, such as cockpit and bridge illumination, observatories, etc. Yellow LED lights are a good choice to meet these special requirements because the human eye is more sensitive to yellow light.

Exercise: The Light Emitting Diode

1. What is an LED?
2. List 5 applications of LEDs.

29.4.3 Transistor

The diode is the simplest semiconductor device, made up of a p-type semiconductor and an n-type semiconductor in contact. It can conduct in only one direction, but it cannot control the size of an electric current. Transistors are more complicated electronic components which can control the size of the electric current flowing through them.

This enables them to be used in amplifiers. A small signal from a microphone or a radio antenna can be used to control the transistor. In response, the transistor will then increase and decrease a much larger current which flows through the speakers.

One of the earliest popular uses of transistors was in cheap and portable radios. Before that, radios were much more expensive and contained glass valves which were fragile and needed replacing. In some parts of the world you can still hear people talking about their 'transistor' — they mean their portable radio.

You can also use a small current to turn the transistor on and off. The transistor then controls a more complicated or powerful current through other components. When a transistor is used in this way it is said to be in **switching** mode as it is acting as a remotely controlled switch. As we shall see in the final sections of this chapter, switching circuits can be used in a computer to process and store digital information. A computer would not work without the millions (or billions) of transistors in it.

There are two main types of transistor - bipolar transistors (NPN or PNP), and field effect transistors (FETs). Both use doped semiconductors, but in different ways. You are mainly required to know about field effect transistors (FETs), however we have to give a brief description of bipolar transistors so that you see the difference.

Bipolar Transistors

Bipolar transistors are made of a doped semiconductor 'sandwich'. In an NPN transistor, a very thin layer of p-type semiconductor is in between two thicker layers of n-type

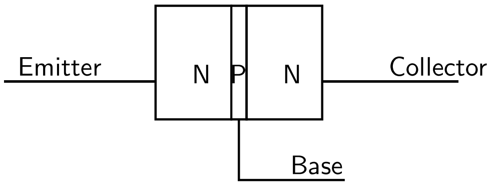


Figure 29.11: An NPN transistor. This is a type of bipolar transistor.

semiconductor. This is shown in Figure 29.11. Similarly an PNP transistor consists of a very thin n-type layer in between two thicker layers of p-type semiconductor.

In an NPN transistor a small current of electrons flows from the emitter (E) to the base (B). Simultaneously, a much larger current of electrons flows from the emitter (E) to the collector (C). If you lower the number of electrons able to leave the transistor at the base (B), the transistor automatically reduces the number of electrons flowing from emitter (E) to collector (C). Similarly, if you increase the current of electrons flowing out of the base (B), the transistor automatically also increases the current of electrons flowing from emitter (E) to collector (C). The transistor is designed so that the current of electrons from emitter to collector (I_{EC}) is proportional to the current of electrons from emitter to base (I_{EB}). The constant of proportionality is known as the **current gain** β. So $I_{EC} = \beta I_{EB}$.

How does it do it? The answer comes from our work with diodes. Electrons arriving at the emitter (n-type semiconductor) will naturally flow through into the central p-type since the base-emitter junction is forward biased. However if none of these electrons are removed from the base, the electrons flowing into the base from the emitter will fill all of the available 'holes'. Accordingly, a large depletion band will be set up. This will act as an insulator preventing current flow into the collector as well. On the other hand, if the base is connected to a positive voltage, a small number of electrons will be removed by the base connection. This will prevent the 'holes' in the base becoming filled up, and no depletion band will form. While some electrons from the emitter leave via the base connection, the bulk of them flow straight on to the collector. You may wonder how the electrons get from the base into the collector (it seems to be reverse biased). The answer is complicated, but the important fact is that the p-type layer is extremely thin. As long as there is no depletion layer, the bulk of the electrons will have no difficulty passing straight from the n-type emitter into the n-type collector. A more satisfactory answer can be given to a university student once band theory has been explained.

Summing up, in an NPN transistor, a small flow of electrons from emitter (E) to base (B) allows a much larger flow of electrons from emitter (E) to collector (C). Given that conventional current (flowing from $+$ to $-$) is in the opposite direction to electron flow, we say that a small conventional current from base to emitter allows a large current to flow from collector to emitter.

A PNP transistor works the other way. A small conventional current from emitter to base allows a much larger conventional current to flow from emitter to collector. The operation is more complicated to explain since the principal charge carrier in a PNP transistor is not the electron but the 'hole'.

The operation of NPN and PNP transistors (in terms of conventional currents) is summarized in Figure 29.12.

Interesting Fact

The transistor is considered by many to be one of the greatest discoveries or inventions in modern history, ranking with banking and the printing press. Key to the importance of the transistor in modern society is its ability to be produced in huge numbers using simple techniques, resulting in vanishingly small prices. Computer "chips" consist of millions of transistors and sell for

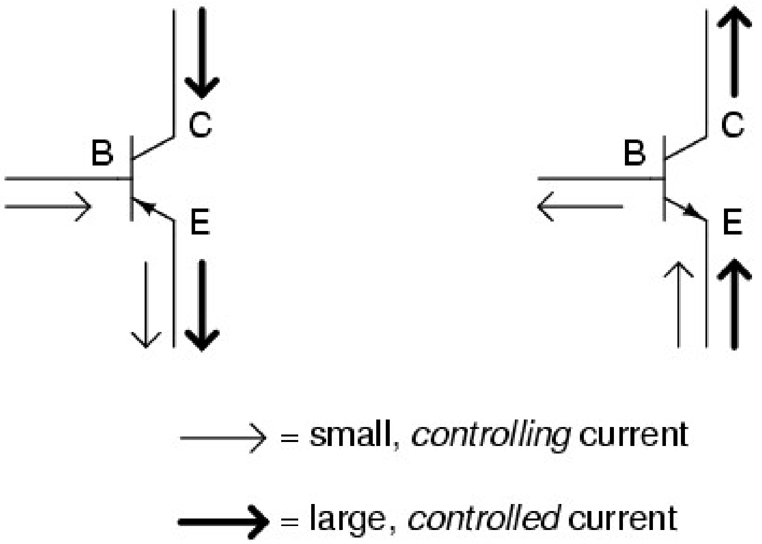


Figure 29.12: An overview of bipolar transistors as current amplifiers.

Rands, with per-transistor costs in the thousandths-of-cents. The low cost has meant that the transistor has become an almost universal tool for non-mechanical tasks. Whereas a common device, say a refrigerator, would have used a mechanical device for control, today it is often less expensive to simply use a few million transistors and the appropriate computer program to carry out the same task through "brute force". Today transistors have replaced almost all electromechanical devices, most simple feedback systems, and appear in huge numbers in everything from computers to cars.

The transistor was invented at Bell Laboratories in December 1947 (first demonstrated on December 23) by John Bardeen, Walter Houser Brattain, and William Bradford Shockley, who were awarded the Nobel Prize in physics in 1956.

The Field Effect Transistor (FET)

To control a bipolar transistors, you control the **current** flowing into or out of its base. The other type of transistor is the field effect transistor (FET). FETs work using control **voltages** instead. Accordingly they can be controlled with much smaller currents and are much more economic to use.

No-one would build a computer with billions of bipolar transistors — the current in each transistor's base might be small, but when you add up all of the base currents in the millions of transistors, the computer as a whole would be consuming a great deal of electricity and making a great deal of heat. Not only is this wasteful, it would prevent manufacturers making a computer of convenient size. If the transistors were too close together, they would overheat.

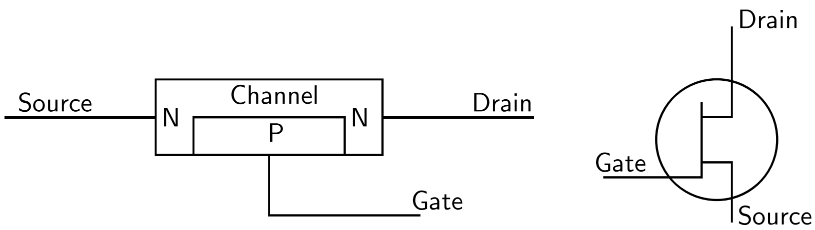


Figure 29.13: A field effect transistor (FET). The diagram on the left shows the semiconductor structure. The diagram on the right shows its circuit symbol.

The three terminals of the FET are called the *source* (S), *drain* (D) and *gate* (G), as shown in Figure 29.13. When the gate is not connected, a current of electrons can flow from source (S) to drain (D) easily along the channel. The source is, accordingly, the negative terminal of the transistor. The drain, where the electrons come out, is the positive terminal of the transistor. A few electrons will flow from the n-type channel into the p-type semiconductor of the gate when the device is manufactured. However, as these electrons are not removed (the gate is not connected), a depletion band is set up which prevents further flow into the gate.

In operation, the gate is connected to negative voltages relative to the source. This makes the p-n junction between gate and channel reverse-biased. Accordingly no current flows from the source into the gate. When the voltage of the gate is lowered (made more negative), the depletion band becomes wider. This enlarged depletion band takes up some of the space of the channel. So the lower the voltage of the gate (the more negative it is relative to the source), the larger the depletion band. The larger the depletion band, the narrower the channel. The narrower the channel, the harder it is for electrons to flow from source to drain.

The voltage of the gate is not the only factor affecting the current of electrons between the source and the drain. If the external circuit has a low resistance, electrons are able to leave the drain easily. If the external circuit has a high resistance, electrons leave the drain slowly. This creates a kind of 'traffic jam' which slows the passage of further electrons. In this way, the voltage of the drain regulates itself, and is more or less independent of the current demanded from the drain.

Once these two factors have been taken into account, it is fair to say that the positive output voltage (the voltage of the drain relative to the source) is proportional to the negative input voltage (the voltage of the gate relative to the source).

For this reason, the field effect transistor is known as a voltage amplifier. This contrasts with the bipolar transistor which is a current amplifier.

?

Exercise: Field Effect Transistors

1. What are the two types of bipolar transistor? How does their construction differ?
2. What are the three connections to a bipolar transistor called?
3. Why are very few electrons able to flow from emitter to collector in an NPN transistor if the base is not connected?

4. Why do you think a bipolar transistor would not work if the base layer were too thick?
5. "The bipolar transistor is a current amplifier." What does this statement mean?
6. Describe the structure of a FET.
7. Define what is meant by the source, drain and gate. During normal operation, what will the voltages of drain and gate be with respect to the source?
8. Describe how a depletion layer forms when the gate voltage is made more negative. What controls the width of the depletion layer?
9. "The field effect transistor is a voltage amplifier." What does this statement mean?
10. The amplifier in a cheap radio will probably contain bipolar transistors. A computer contains many field effect transistors. Bipolar transistors are more rugged and less sensitive to interference than field effect transistors, which makes them more suitable for a simple radio. Why are FETs preferred for the computer?

29.4.4 The Operational Amplifier

The operational amplifier is a special kind of voltage amplifier which is made from a handful of bipolar or field effect transistors. Operational amplifiers are usually called **op-amps** for short. They are used extensively in all kinds of audio equipment (amplifiers, mixers and so on) and in instrumentation. They also have many other uses in other circuit - for example comparing voltages from sensors.

Operational amplifiers are supplied on Integrated Circuits (I.C.s). The most famous operational amplifier I.C. is numbered 741 and contains a single operational amplifier on an integrated circuit ('chip') with eight terminals. Other varieties can be bought, and you can get a single integrated circuit with two or four '741'-type operational amplifiers on it.

The symbol for an op-amp is shown in Figure 29.14. The operational amplifier has two input terminals and one output terminal. The voltage of the output terminal is proportional to the difference in voltage between the two input terminals. The output terminal is on the right (at the sharp point of the triangle). The two input terminals are drawn on the left. One input terminal (labelled with a $+$ on diagrams) is called the **non-inverting input**. The other input terminal (labelled $-$) is called the **inverting input**. The labels $+$ and $-$ have nothing to do with the way in which the operational amplifier is connected to the power supply. Operational amplifiers must be connected to the power supply, but this is taken for granted when circuit diagrams are drawn, and these connections are not shown on circuit diagrams. Usually, when drawing electronic circuits, '0V' is taken to mean the negative terminal of the power supply. This is not the case with op-amps. For an op-amp, '0V' refers to the voltage midway between the $+$ and $-$ of the supply.

The output voltage of the amplifier V_{out} is given by the formula

$$V_{out} = A\left(V_{+} - V_{-}\right) \qquad (29.5)$$

here A is a constant called the **open loop gain**, and V_{+} and V_{-} are the voltages of the two input terminals. That said, the output voltage can not be less than the voltage of the negative terminal of the battery supplying it or higher than the positive terminal of the battery supplying it. You will notice that V_{out} is positive if $V_{+} > V_{-}$ and negative if $V_{+} < V_{-}$. This is why the $-$ input is called the inverting input: raising its voltage causes the output voltage to *drop*.

The input resistance of an operational amplifier is very high. This means that very little current flows into the input terminals during operation.

If all of the transistors in the operational amplifier were identical then the output voltage would be zero if the two inputs were at equal voltages. In practice this is not quite the case, and for sensitive work a **trimming potentiometer** is connected. This is adjusted until the op-amp is zeroed correctly.

Simple operational amplifiers require the trimming potentiometer to be built into the circuit containing them, and an example is shown in Figure 29.15. Other operational amplifier designs incorporate separate terminals for the trimming potentiometer. These special terminals are labelled **offset** on the manufacturer's diagram. The exact method of connecting the potentiometer to the offset terminals can depend on the design of the operational amplifier, and you need to refer to the manufacturer's data sheet for details of which potentiometer to use and how to connect it.

For most commercially produced operational amplifiers (known as op-amps for short), the open loop gain A is very large and does not stay constant. Values of 100 000 are typical. Usually a designer would want an amplifier with a stable gain of smaller value, and builds the operational amplifier into a circuit like the one in Figure 29.15.

Extension: Calculating the gain of the amplifier in Figure 29.15.

1. The input resistance of the operational amplifier is very high. This means that very little current flows into the inverting input of the op-amp. Accordingly, the current through resistor R_1 must be almost the same as the current through resistor R_2. This means that the ratio of the voltage across R_1 to the voltage across R_2 is the same as the ratio of the two resistances.
2. The open loop gain A of the op-amp is very high. Assuming that the output voltage is less than a few volts, this means that the two input terminals must be at very similar voltages. We shall assume that they are at the same voltage.
3. We want the output voltage to be zero if the input voltage is zero. Assuming that the transistors within the op-amp are very similar, the output voltage will only be zero for zero input voltage if V_+ is very close to zero. We shall assume that $V_+ = 0$ when the trimming potentiometer is correctly adjusted.
4. It follows from the last two statements that $V_- \approx 0$, and we shall assume that it is zero.
5. With these assumptions, the voltage across R_2 is the same as V_{out}, and the voltage across R_1 is the same as V_{in}. Since both resistors carry the same current (as noted in point 1), we may say that the magnitude of $V_{out}/V_{in} = R_2/R_1$. However, if V_{in} is negative, then V_{out} will be positive. Therefore it is customary to write the gain of this circuit as $V_{out}/V_{in} = -R_2/R_1$.

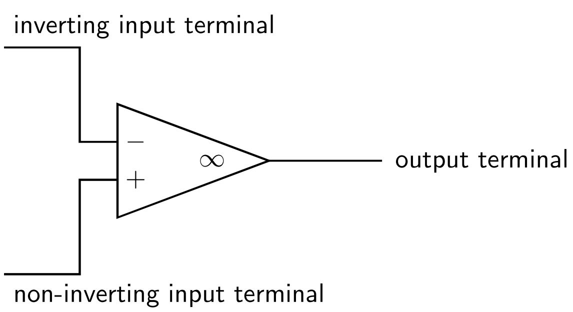


Figure 29.14: Circuit symbol for an operational amplifier. The amplifier must also be connectd to the $+$ and $-$ terminals of the power supply. These connections are taken for granted and not shown.

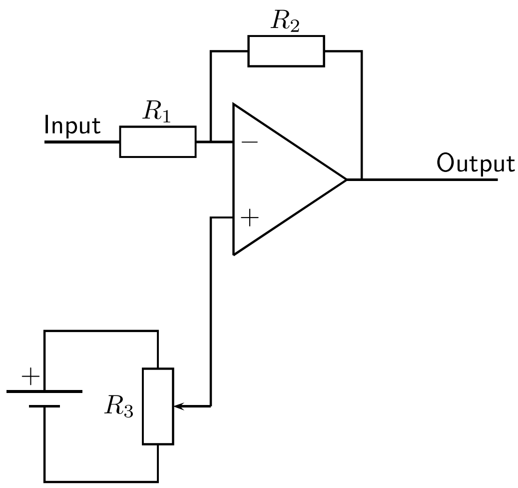


Figure 29.15: An inverting amplifier built using an operational amplifier. The connections from battery to operational amplifier are not shown. The output voltage $V_{out} = -R_2V_{in}/R_1$, as explained in the text. The potentiometer R_3 is a trimming potentiometer. To set it, the input is connected to zero volts. The trimming potentiometer is then adjusted until $V_{out} = 0$. In all operational amplifier circuits, zero volts is midway between the + and − of the supply.

?

Exercise: Operational Amplifiers

1. What are operational amplifiers used for?
2. Draw a simple diagram of an operational amplifier and label its terminals.
3. Why is a trimming potentiometer is needed when using an op-amp?

29.5 The Principles of Digital Electronics

The circuits and components we have discussed are very useful. You can build a radio or television with them. You can make a telephone. Even if that was all there was to electronics, it would still be very useful. However, the great breakthrough in the last fifty years or so has been in **digital** electronics. This is the subject which gave us the computer. The computer has revolutionized the way business, engineering and science are done. Small computers programmed to do a specific job (called microprocessors) are now used in almost every electronic machine from cars to washing machines. Computers have also changed the way we communicate. We used to have telegraph or telephone wires passing up and down a country — each one carrying one telephone call or signal. We now have optic fibres each capable of carrying tens of thousands of telephone calls using **digital** signals.

So, what is a digital signal? Look at Figure 29.16. A normal signal, called an **analogue** signal, carries a smooth wave. At any time, the voltage of the signal could take any value. It could be 2,00 V or 3,53 V or anything else. A digital signal can only take certain voltages. The simplest case is shown in the figure — the voltage takes one of two values. It is either **high**, or it is **low**. It never has any other value.

These two special voltages are given symbols. The low voltage level is written 0, while the high voltage level is written as 1. When you send a digital signal, you set the voltage you want (0 or 1), then keep this fixed for a fixed amount of time (for example 0.01 μs), then you send the next 0 or 1. The digital signal in Figure 29.16 could be written 01100101.

Why are digital signals so good?

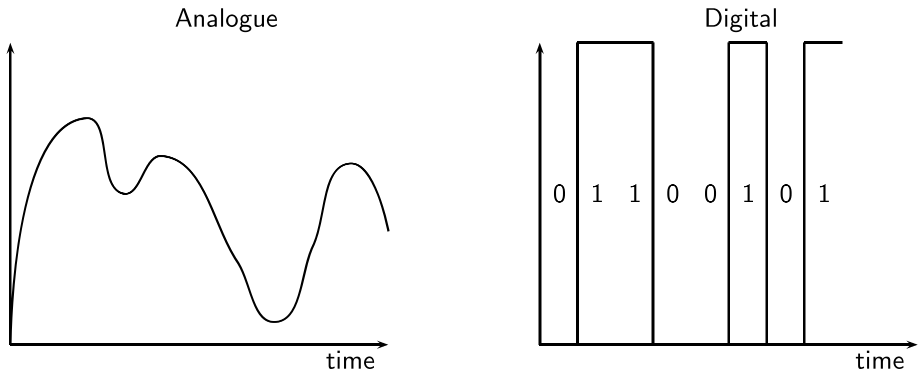


Figure 29.16: The difference between normal (analogue) signals and digital signals. The analogue signal is on the left.

1. Using a computer, any information can be turned into a pattern of 0s and 1s. Pictures, recorded music, text and motion pictures can all be turned into a string of 0s and 1s and transmitted or stored in the same way. The computer receiving the signal at the other end converts it back again. A Compact Disc (CD) for example, can store music or text or pictures, and all can be read using a computer.

2. The 0 and the 1 look very different. You can immediately tell if a 0 or a 1 is being sent. Even if there is interference, you can still tell whether the sender sent a 0 or a 1. This means that fewer mistakes are made when reading a digital signal. This is why the best music recording technologies, and the most modern cameras, for example, all use digital technology.

3. Using the 0s and 1s you can count, and do all kinds of mathematics. This will be explained in more detail in the next section.

The simplest digital circuits are called **logic gates**. Each logic gate makes a decision based on the information it receives. Different logic gates are set up to make the decisions in different ways. Each logic gate will be made of many microscopic transistors connected together within a thin wafer of silicon. This tiny circuit is called an Integrated Circuit or I.C. - all the parts are in one place (integrated) on the silicon wafer.

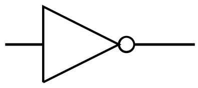
29.5.1 Logic Gates

There are five main types of logic gate: NOT, AND, OR, NAND and NOR. Each one makes its decision in a different way.

The NOT Gate

Problem: You want an automatic circuit in your office to turn on the heating in the winter. You already have a digital electronic temperature sensor. When the temperature is high, it sends out a 1. When the office is cold, it sends out a 0. If this signal were sent straight to the heater, the heater would turn on (1) when it was already hot, and would stay off when it was cold. This is wrong! To make the heater work, we need a circuit which will change a 0 (from the sensor) into a 1 (to send to the heater). This will make the heater come on when it is cold. You also want it to change a 1 (from the sensor) into a 0 (to send to the heater). This will turn the heater off when the room is hot. This circuit is called an **inverter** or **NOT gate**. It changes 0 into 1 (1 is NOT 0). It changes 1 into 0 (0 is NOT 1). It changes a signal into what it is NOT.

The symbol for the NOT gate is:



The action of the NOT gate is written in a table called a **truth table**. The left column shows the possible inputs on different rows. The right column shows what the output (decision) of the circuit will be for that input. The truth table for the NOT gate is shown below.

Input	Output
0	1
1	0

When you read the truth table, the top row says, "If the input is 0, the output will be 1." For our heater, this means, "If the room is cold, the heater will turn on." The bottom row says, "If the input is 1, the output will be 0." For our heater, this means, "If the room is hot, the heater will switch off."

The AND Gate

Problem: An airliner has two toilets. Passengers get annoyed if they get up from their seat only to find that both toilets are being used and they have to go back to their seat and wait. You want to fit an automatic circuit to light up a display if both toilets are in use. Then passengers know that if the light is off, there will be a free toilet for them to use. There is a sensor in each toilet. It gives out a 0 of the toilet is free, and a 1 if it is in use. You want to send a 1 to the display unit if **both** sensors are sending 1s. To do this, you use an AND gate.

The symbol for the AND gate is:

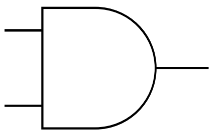


Figure 29.17: Symbol for the AND logic gate.

The truth table for the AND gate is shown below. An AND gate has two inputs (the NOT gate only had one). This means we need four rows in the truth table, one for each possible set of inputs. The first row, for example, tells us what the AND gate will do if both inputs are 0. In our airliner, this means that both toilets are free. The right column has a 0 showing that the output will be 0, so the display will not light up. The second row has inputs of 0 and 1 (the first toilet is free, the other is in use). Again the output is 0. The third row tells us what will happen if the inputs are 1 and 0 (the first toilet is in use, and the second is free). Finally, the last line tells us what will happen if both inputs are 1 (both toilets are in use). It is only in this case that the output is 1 and the display lights up.

Inputs		Output
A	B	
0	0	0
0	1	0
1	0	0
1	1	1

This device is called an AND gate, because the output is only 1 if one input AND the other input are both 1.

Extension: Using 0 and 1 to mean True and False

When we use logic gates we use the low voltage state 0 to represent 'false'. The high voltage state 1 represents 'true'. This is why the word AND is so appropriate. A AND B is true (1) if, and only if, A is true (1) AND B is true (1).

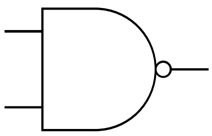
Extension: AND and multiplication

Sometimes, the AND operation is written as multiplication. A AND B is written AB. If either A or B are 0, then AB will also be 0. For AB to be 1, we need A and B to both be 1. Multiplication of the numbers 0 and 1 does exactly the same job as an AND gate.

The NAND Gate

Problem: You build the circuit for the airliner toilets using an AND gate. Your customer is pleased, but she says that it would be better if the display lit up when there **was** a free toilet. In other words, the display should light up unless both toilets are in use. To do this we want a circuit which does the opposite of an AND gate. We want a circuit which would give the output 0 if an AND gate would give 1. We want a circuit which would give the output 1 if an AND gate would give 0. This circuit is called a NAND gate.

The symbol for the NAND gate is:



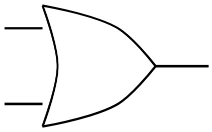
The truth table for the NAND gate is shown below.

Inputs		Output
A	B	
0	0	1
0	1	1
1	0	1
1	1	0

You may have noticed that we could have done this job on the airliner by using our earlier circuit, with a NOT gate added between the original AND gate and the display. This is where the word NAND comes from — it is short for NotAND.

The OR Gate

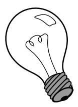
Problem: A long, dark corridor has two light switches — one at each end of the corridor. The switches each send an output of 0 to the control unit if no-one has pressed the switch. If someone presses the switch, its output is 1. The lights in the corridor should come on if either switch is pressed. To do this job, the control unit needs an OR gate. The symbol for the OR gate is:



The truth table for the OR gate is shown.

Inputs		Output
A	B	
0	0	0
0	1	1
1	0	1
1	1	1

You can see that the output is 1 (and the lights come on in the corridor) if either one switch OR the other is pressed. Pressing both switches also turns on the lights, as the last row in the table shows.



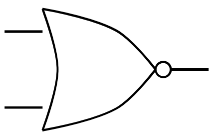
Extension: OR and addition

Sometimes you will see A OR B written mathematically as A+B. This makes sense, since if A=0 and B=0, then A OR B = A+B = 0. Similarly, if A=0 and B=1, then A OR B = A+B = 1. If A=1 and B=0, then A OR B = A+B = 1 once again. The only case where the OR function differs from normal addition is when A=1 and B=1. Here A OR B = 1 in logic, but A+B=2 in arithmetic. However, there is no such thing as '2' in logic, so we define + to mean 'OR', and write 1+1=1 with impunity!

If you wish, you can prove that the normal rules of algebra still work using this notation: A+(B+C) = (A+B)+C, A(BC) = (AB)C, and A(B+C) = AB + AC. This special kind of algebra where variables can only be 0 (representing false) or 1 (representing true) is called Boolean algebra.

The NOR Gate

The last gate you need to know is the NOR gate. This is opposite to the OR gate. The output is 1 if both inputs are 0. In other words, the output switches on if neither the first NOR the second input is 1. The symbol for the NOR gate is:



The truth table for the NOR gate is shown below.

Inputs		Output
A	B	
0	0	1
0	1	0
1	0	0
1	1	0

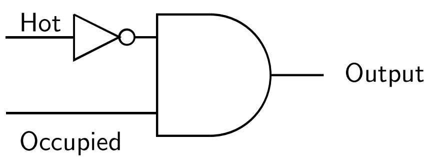
The examples given were easy. Each job only needed one logic gate. However any 'decision making' circuit can be built with logic gates, no matter how complicated the decision. Here is an example.

Worked Example 181: An Economic Heating Control

Question: A sensor in a building detects whether a room is being used. If it is empty, the output is 0, if it is in use, the output is 1. Another sensor measures the temperature of the room. If it is cold, the output is 0. If it is hot, the output is 1. The heating comes on if it receives a 1. Design a control circuit so that the heating only comes on if the room is in use and it is cold.

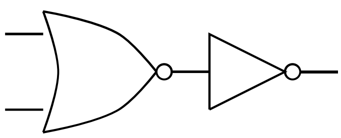
Answer

The first sensor tells us whether the room is occupied. The second sensor tells us whether the room is hot. The heating must come on if the room is occupied AND cold. This means that the heating should come on if the room is occupied AND (NOT hot). To build the circuit, we first attach a NOT gate to the output of the temperature sensor. This output of the NOT gate will be 1 only if the room is cold. We then attach this output to an AND gate, together with the output from the other sensor. The output of the AND gate will only be 1 if the room is occupied AND the output of the NOT gate is also 1. So the heating will only come on if the room is in use and is cold. The circuit is shown below.



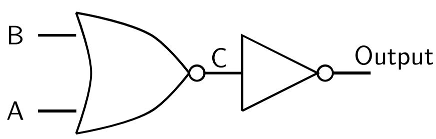
Worked Example 182: Solving a circuit with two logic gates

Question: Compile the truth table for the circuit below.



Answer

Firstly, we label the inputs A and B. We also label the point where the two gates are connected C.



Next we prepare a truth table. There is a column for each of the inputs, for the intermediate point C and also for the output. The truth table has four rows, since there are four possible inputs — 00, 01, 10 and 11.

A	B	C	Output
0	0		
0	1		
1	0		
1	1		

Next we fill in the C column given that we know what a NOR gate does.

A	B	C	Output
0	0	1	
0	1	0	
1	0	0	
1	1	0	

Next, we can fill in the output, since it will always be the opposite of C (because of the NOT gate).

A	B	C	Output
0	0	1	0
0	1	0	1
1	0	0	1
1	1	0	1

Finally we see that this combination of gates does the same job as an OR gate.

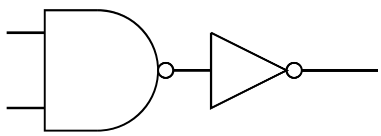
Each logic gate is manufactured from two or more transistors. Other circuits can be made using logic gates, as we shall see in the next section. We shall show you how to count and store numbers using logic gates. This means that if you have enough transistors, and you connect them correctly to make the right logic gates, you can make circuits which count and store numbers.

In practice, the cheapest gate to manufacture is usually the NAND gate. Additionally, Charles Peirce showed that NAND gates alone (as well as NOR gates alone) can be used to reproduce all the other logic gates.

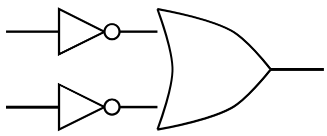
?

Exercise: The Principles of Digital Electronics

1. Why is digital electronics important to modern technology and information processing?
2. What two symbols are used in digital electronics, to represent a "high" and a "low"? What is this system known as?
3. What is a logic gate?
4. What are the five main types of logic gates? Draw the symbol for each logic gate.
5. Write out the truth tables for each of the five logic gates.
6. Write out the truth table for the following circuit. Which single gate is this circuit equivalent to?



7. Write out the truth table for the following circuit. Which single gate is this circuit equivalent to?



29.6 Using and Storing Binary Numbers

In the previous section, we saw how the numbers 0 and 1 could represent 'false' and 'true' and could be used in decision making. Often we want to program a computer to count with numbers. To do this we need a way of writing any number using nothing other than 0 and 1. When written in this way, numbers are called **binary numbers**.

Definition: Binary
A way of writing any number using only the digits 0 and 1.

29.6.1 Binary numbers

In normal (denary) numbers, we write 9+1 as 10. The fact that the '1' in 10 is the second digit from the right tells us that it actually means 10 and not 1. Similarly, the '3' in 365 represents 300 because it is the third digit from the right. You could write 365 as $3 \times 100 + 6 \times 10 + 5$. You will notice the pattern that the nth digit from the right represents 10^{n-1}. In binary, we use the nth digit from the right to represent 2^{n-1}. Thus 2 is written as 10 in binary. Similarly $2^2 = 4$ is written as 100 in binary, and $2^3 = 8$ is written as 1000 in binary.

Worked Example 183: Conversion of Binary Numbers to Denary Numbers

Question: Convert the binary number 10101 to its denary equivalent.
Answer
We start on the right. The '1' on the right does indeed represent one. The next '1' is in the third place from the right, and represents $2^2 = 4$. The next '1' is in the fifth place from the right and represents $2^4 = 16$. Accordingly, the binary number 10101 represents 16+4+1 = 21 in denary notation.

Worked Example 184: Conversion of Denary Numbers to Binary Numbers

Question: Convert the decimal number 12 to its binary equivalent.
Answer
Firstly we write 12 as a sum of powers of 2, so 12 = 8+4. In binary, eight is 1000, and four is 100. This means that twelve = eight + four must be 1000+100 = 1100 in binary. You could also write 12 as $1 \times 8 + 1 \times 4 + 0 \times 2 + 0 \times 1 = 1100$ in binary.

Interesting Fact

How do you write numbers as a sum of powers of two? The first power of two (the largest) is the largest power of two which is not larger than the number you are working with. In our last example, where we wanted to know what twelve was in binary, the largest power of two which is not larger than 12 is 8. Thus 12 = 8 + something. By arithmetic, the 'something' must be 4, and the largest power of two not larger than this is 4 exactly. Thus 12 = 8 + 4, and we have finished.

A more complicated example would be to write one hundred in binary. The largest power of two not larger than 100 is 64 (1000000 in binary). Subtracting 64 from 100 leaves 36. The largest power of two not larger than 36 is 32 (100000 in binary). Removing this leaves a remainder of 4, which is a power of two itself (100 in binary). Thus one hundred is 64 + 32 + 4, or in binary 1000000 + 100000 + 100 = 1100100.

Once a number is written in binary, it can be represented using the low and high voltage levels of digital electronics. We demonstrate how this is done by showing you how an electronic counter works.

29.6.2 Counting circuits

To make a counter you need several 'T flip flops', sometimes called 'divide by two' circuits. A T flip flop is a digital circuit which swaps its output (from 0 to 1 or from 1 to 0) whenever the input changes from 1 to 0. When the input changes from 0 to 1 it doesn't change its output. It is called a **flip flop** because it changes (flips or flops) each time it receives a pulse.

If you put a series of pulses 10101010 into a T flip flop, the result is 01100110. Figure 29.18 makes this clearer.

As you can see from Figure 29.18, there are half as many pulses in the output. This is why it is called a 'divide by two' circuit.

If we connect T flip flops in a chain, then we make a counter which can count pulses. As an example, we connect three T flip flops in a chain. This is shown in Figure 29.19.

When this circuit is fed with a stream of pulses, the outputs of the different stages change. The table below shows how this happens. Each row shows a different stage, with the first stage at the top. We assume that all of the flip flops have 0 as their output to start with.

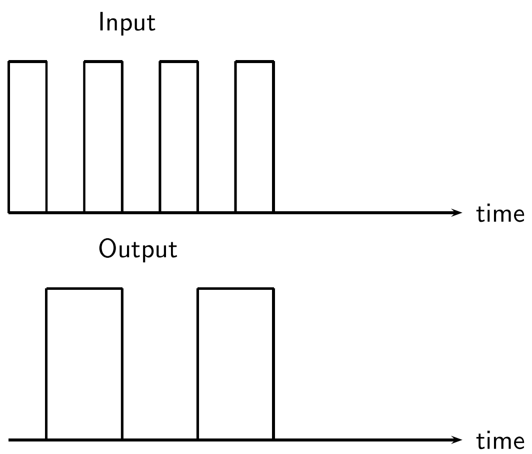


Figure 29.18: The output of a T flip flop, or 'divide by two' circuit when a square wave is connected to the input. The output changes state when the input goes from 1 to 0.

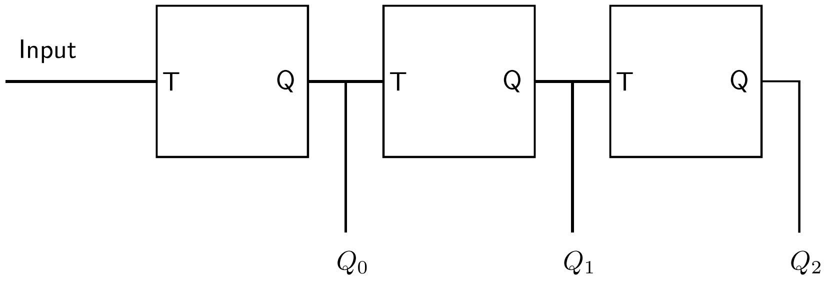


Figure 29.19: Three T flip flops connected together in a chain to make a counter. The input of each flip flop is labelled T, while each output is labelled Q. The pulses are connected to the input on the left. The outputs Q_0, Q_1 and Q_2 give the three digits of the binary number as the pulses are counted. This is explained in the text and in the next table.

Input	Output 1	Output 2	Output 3	Number of pulse	Number in binary
1	0	0	0	0	000
0	1	0	0	1	001
1	1	0	0	1	001
0	0	1	0	2	010
1	0	1	0	2	010
0	1	1	0	3	011
1	1	1	0	3	011
0	0	0	1	4	100
1	0	0	1	4	100
0	1	0	1	5	101
1	1	0	1	5	101
0	0	1	1	6	110
1	0	1	1	6	110
0	1	1	1	7	111
1	1	1	1	7	111
0	0	0	0	8	1000
1	0	0	0	8	1000
0	1	0	0	9	1101
1	1	0	0	9	1101

The binary numbers in the right hand column count the pulses arriving at the input. You will notice that the output of the first flip flop gives the right most digit of the pulse count (in binary). The output of the second flip flop gives the second digit from the right (the 'twos' digit) of the pulse count. The output of the third flip flop gives the third digit from the right (the 'fours' digit) of the pulse count. As there are only three flip flops, there is nothing to provide the next digit (the 'eights' digit), and so the eighth pulse is recorded as 000, not 1000.

This device is called a **modulo 8** counter because it can count in eight stages from 000 to 111 before it goes back to 000. If you put four flip flops in the counter, it will count in sixteen stages from 0000 to 1111, and it is called a modulo 16 counter because it counts in sixteen stages before going back to 0000.

Definition: Modulo
The modulo of a counter tells you how many stages (or pulses) it receives before going back to 0 as its output. Thus a modulo 8 counter counts in eight stages 000, 001, 010, 011, 100, 101, 110, 111, then returns to 000 again.

Interesting Fact

If a counter contains n flip flops, it will be a modulo 2^n counter. It will count from 0 to $2^n - 1$.

29.6.3 Storing binary numbers

Counting is important. However, it is equally important to be able to remember the numbers. Computers can convert almost anything to a string of 0s and 1s, and therefore to a binary number. Unless this number can be stored in the computer's memory, the computer would be useless.

The memory in the computer contains many parts. Each part is able to store a single 0 or 1. Since 0 and 1 are the two binary digits, we say that each part of the memory stores one **bit**.

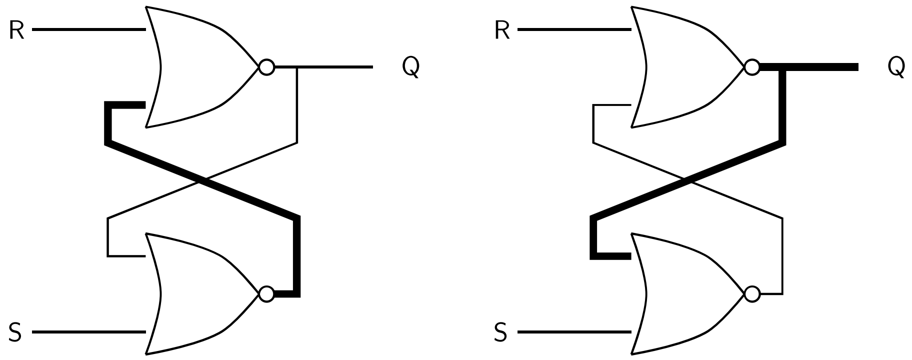


Figure 29.20: A bistable circuit made from two NOR gates. This circuit is able to store one bit of digital information. With the two inputs set to 0, you can see that the output could be (and will remain) either 0 or 1. The circuit on the left shows an output of 0, the circuit on the right shows an output of 1. Wires carrying high logic levels (1) are drawn thicker. The output of the bistable is labelled Q.

Definition: Bit
One bit is a short way of saying one 'binary digit'. It is a single 0 or 1.

If you have eight bits, you can store a binary number from 00000000 to 11111111 (0 to 255 in denary). This gives you enough permutations of 0s and 1s to have one for each letter of the alphabet (in upper and lower case), each digit from 0 to 9, each punctuation mark and each control code used by a computer in storing a document. When you type text into a word processor, each character is stored as a set of eight bits. Each set of eight bits is called a **byte**. Computer memories are graded according to how many bytes they store. There are 1024 bytes in a kilobyte (kB), 1024×1024 bytes in a megabyte (MB), and $1024 \times 1024 \times 1024$ bytes in a gigabyte (GB).

To store a bit we need a circuit which can 'remember' a 0 or a 1. This is called a **bistable** circuit because it has two stable states. It can stay indefinitely either as a 0 or a 1. An example of a bistable circuit is shown in Figure 29.20. It is made from two NOR gates.

To store the 0 or the 1 in the bistable circuit, you set one of the inputs to 1, then put it back to 0 again. If the input labelled 'S' (set) is raised, the output will immediately become 1. This is shown in Figure 29.21.

To store a 0, you raise the 'R' (reset) input to 1. This is shown in Figure 29.22.

Once you have used the S or R inputs to set or reset the bistable circuit, you then bring both inputs back to 0. The bistable 'remembers' the state. Because of the ease with which the circuit can be Reset and Set it is also called a **RS flip flop** circuit.

A computer memory will be able to store millions or billions of bits. If it used our circuit above, it would need millions or billions of NOR gates, each of which is made from several transistors. The computer memory is made of many millions of transistors.

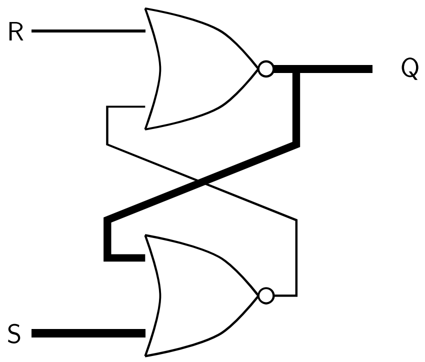


Figure 29.21: The output of a bistable circuit is **set** (made 1) by raising the 'S' input to 1. Wires carrying high logic levels (1) are shown with thicker lines.

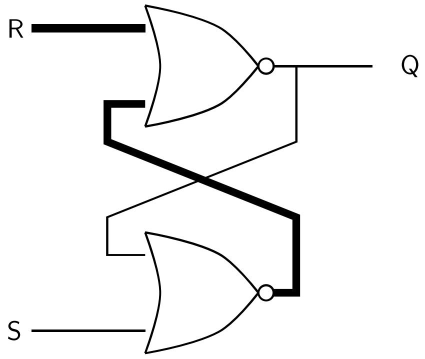


Figure 29.22: The output of a bistable circuit is **reset** (made 0) by raising the 'R' input to 1. Wires carrying high logic levels (1) are shown with thicker lines.

The bistable circuits drawn here don't remember 0s or 1s for ever — they lose the information if the power is turned off. The same is true for the RAM (Random Access Memory) used to store working and temporary data in a computer. Some modern circuits contain special memory which can remember its state even if the power is turned off. This is used in FLASH drives, commonly found in USB data sticks and on the memory cards used with digital cameras. These bistable circuits are much more complex.

You can also make T flip flops out of logic gates, however these are more complicated to design.

Exercise: Counting Circuits

1. What is the term *bit* short for?
2. What is 43 in binary?
3. What is 1100101 in denary?
4. What is the highest number a modulo 64 counter can count to? How many T flip flops does it contain?
5. What is the difference between an RS flip flop and a T flip flop?
6. Draw a circuit diagram for a bistable circuit (RS flip flop). Make three extra copies of your diagram. On the first diagram, colour in the wires which will carry high voltage levels (digital 1) if the R input is low, and the S input is high. On the second diagram, colour in the wires which carry high voltage levels if the S input of the first circuit is now made low. On the third diagram, colour in the wires which carry high voltage levels if the R input is now made high. On the final diagram, colour in the wires carrying high voltage levels if the R input is now made low again.
7. Justify the statement: a modern computer contains millions of transistors.

Exercise: End of Chapter Exercises

1. Calculate the reactance of a 3 mH inductor at a frequency of 50 Hz.
2. Calculate the reactance of a 30 μF capacitor at a frequency of 1 kHz.
3. Calculate the impedance of a series circuit containing a 5 mH inductor, a 400 μF capacitor and a 2 kΩ resistor at a frequency of 50 kHz.
4. Calculate the frequency at which the impedance of the circuit in the last question will be the smallest.
5. Which component can be used to block low frequencies?
6. Draw a circuit diagram with a battery, diode and resistor in series. Make sure that the diode is forward biased so that a current will flow through it.

7. When building a complex electronic circuit which is going to be powered by a battery, it is always a good idea to put a diode in series with the battery. Explain how this will protect the circuit if the user puts the battery in the wrong way round.
8. Summarize the differences betwen a bipolar and field effect transistor.
9. What does an operational amplifier (op-amp) do?
10. What is the difference between a digital signal and an analogue signal?
11. What are the advantages of digital signals over analogue signals?
12. Draw the symbols for the five logic gates, and write down their truth tables.
13. Draw a circuit diagram with an AND gate. Each input should be connected to the output of a separate NOT gate. By writing truth tables show that this whole circuit behaves as a NOR gate.
14. Convert the denary number 99 into binary.
15. Convert the binary number 11100111 into denary.
16. Explain how three T flip flops can be connected together to make a modulo 8 counter. What is the highest number it can count up to?
17. Draw the circuit diagram for an RS flip flop (bistable) using two NOR gates.
18. Show how the circuit you have just drawn can have a stable output of 0 or 1 when both inputs are 0.
19. Operational (and other) amplifiers, logic gates, and flip flops all contain transistors, and would not work without them. Write a short newspaper article for an intelligent reader who knows nothing about electronics. Explain how important transistors are in modern society.