The Free High School Science Texts: A Textbook for High School Students Studying Physics.

FHSST Authors[1]

December 9, 2005

[1] See http://savannah.nongnu.org/projects/fhsst

Contents

Part I

Physics

Physics is the study of the world around us. In a sense we are more qualified to do physics than any other science. From the day we are born we study the things around us in an effort to understand how they work and relate to each other. Learning how to catch or throw a ball is a physics undertaking for example.

In the field of study we refer to as physics we just try to make the things everyone has been studying more clear. We attempt to describe them through simple rules and mathematics. Mathematics is merely the language we use.

The best approach to physics is to relate everything you learn to things you have already noticed in your everyday life. Sometimes when you look at things closely you discover things you had overlooked intially.

It is the continued scrutiny of everything we know about the world around us that leads people to the lifelong study of physics. You can start with asking a simple question like ”Why is the sky blue?” which could lead you to electromagnetic waves which in turn could lead you wave particle duality and to energy levels of atoms and before long you are studying quantum mechanics or the structure of the universe.

In the sections that follow notice that we will try to describe how we will communicate the things we are dealing with. This is our langauge. Once this is done we can begin the adventure of looking more closely at the world we live in.

Chapter 1

Units

1.1 PGCE Comments

- Explain what is meant by 'physical quantity'.
- Chapter is too full of tables and words; need figures to make it more interesting.
- Make researching history of SI units a small project.
- Multiply by one technique: not positive! Suggest using exponents instead (i.e. use the table of prefixes). This also works better for changing complicated units (km/h^{-1} to $m.s^{-1}$ etc....). Opinion that this technique is limited in its application.
- Edit NASA story.
- The Temperature section should be cut-down. SW: I have edited the original section but perhaps a more aggressive edit is justified with the details deffered until the section on gases.

1.2 'TO DO' LIST

- Write section on scientific notation, significant figures and rounding.
- Add to sanity test table of sensible values for things.
- Graph Celsius/Kelvin ladder.
- Address PGCE comments above.

1.3 Introduction

Imagine you had to make curtains and needed to buy material. The shop assistant would need to know how much material was required. Telling her you need material 2 wide and 6 long would be insufficient— you have to specify the **unit** (i.e. 2 *metres* wide and 6 *metres* long). Without the unit the information is incomplete and the shop assistant would have to guess. If you were making curtains for a doll's house the dimensions might be 2 centimetres wide and 6 centimetres long!

Base quantity	**Name**	**Symbol**
length	metre	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
thermodynamic temperature	kelvin	K
amount of substance	mole	mol
luminous intensity	candela	cd

Table 1.1: SI Base Units

It is not just lengths that have units, all physical quantities have units (e.g. time and temperature).

1.4 Unit Systems

There are many unit systems in use today. Physicists, for example, use 4 main sets of units: SI units, c.g.s units, imperial units and natural units.

Depending on where you are in the world or what area of physics you work in, the units will be different. For example, in South Africa road distances are measured in kilometres (SI units), while in England they are measured in miles (imperial units). You could even make up your own system of units if you wished, but you would then have to teach people how to use it!

1.4.1 SI Units (*Système International d'Unités*)

These are the internationally agreed upon units and the ones we will use. Historically these units are based on the metric system which was developed in France at the time of the French Revolution.

All physical quantities have units which can be built from the 7 base units listed in Table 1.1 (incidentally the choice of these seven was arbitrary). They are called base units because none of them can be expressed as combinations of the other six. This is similar to breaking a language down into a set of sounds from which all words are made. Another way of viewing the base units is like the three primary colours. All other colours can be made from the primary colours but no primary colour can be made by combining the other two primaries.

Unit names are always written with lowercase initials (e.g. the metre). The symbols (or abbreviations) of units are also written with lowercase initials except if they are named after scientists (e.g. the kelvin (K) and the ampere (A)).

To make life convenient, particular combinations of the base units are given special names. This makes working with them easier, but it is always correct to reduce everything to the base units. Table 1.2 lists some examples of combinations of SI base units assigned special names. Do not be concerned if the formulae look unfamiliar at this stage– we will deal with each in detail in the chapters ahead (as well as many others)!

It is very important that you are able to say the units correctly. For instance, the **newton** is another name for the **kilogram metre per second squared** ($kg.m.s^{-2}$), while the **kilogram metre squared per second squared** ($kg.m^2.s^{-2}$) is called the **joule**.

Another important aspect of dealing with units is the prefixes that they sometimes have (prefixes are words or letters written in front that change the meaning). The kilogram (kg) is a simple example. $1kg$ is $1000g$ or $1 \times \mathbf{10^3g}$. Grouping the 10^3 and the g together we can replace

Quantity	**Formula**	**Unit Expressed in Base Units**	**Name of Combination**
Force	ma	$kg.m.s^{-2}$	N (newton)
Frequency	$\frac{1}{T}$	s^{-1}	Hz (hertz)
Work & Energy	$F.s$	$kg.m^2.s^{-2}$	J (joule)

Table 1.2: Some Examples of Combinations of SI Base Units Assigned Special Names

the 10^3 with the prefix k (kilo). Therefore the k takes the place of the 10^3. Incidentally the kilogram is unique in that it is the only SI base unit containing a prefix

There are prefixes for many powers of 10 (Table 1.3 lists a large set of these prefixes). This is a larger set than you will need but it serves as a good reference. The case of the prefix symbol is very important. Where a letter features twice in the table, it is written in uppercase for exponents bigger than one and in lowercase for exponents less than one. **Those prefixes listed in boldface should be learnt.**

Prefix	**Symbol**	**Exponent**	**Prefix**	**Symbol**	**Exponent**
yotta	Y	10^{24}	yocto	y	10^{-24}
zetta	Z	10^{21}	zepto	z	10^{-21}
exa	E	10^{18}	atto	a	10^{-18}
peta	P	10^{15}	femto	f	10^{-15}
tera	T	10^{12}	pico	p	10^{-12}
giga	G	10^{9}	**nano**	n	10^{-9}
mega	M	10^{6}	**micro**	μ	10^{-6}
kilo	k	10^{3}	**milli**	m	10^{-3}
hecto	h	10^{2}	**centi**	c	10^{-2}
deca	da	10^{1}	**deci**	d	10^{-1}

Table 1.3: Unit Prefixes

As another example of the use of prefixes, $1 \times 10^{-3} g$ can be written as $1mg$ (1 milligram).

1.4.2 The Other Systems of Units

The remaining sets of units, although not used by us, are also internationally recognised and still in use by others. We will mention them briefly for interest only.

c.g.s Units

In this system the metre is replaced by the centimetre and the kilogram is replaced by the gram. This is a simple change but it means that all units derived from these two are changed. For example, the units of force and work are different. These units are used most often in astrophysics and atomic physics.

Imperial Units

These units (as their name suggests) stem from the days when monarchs decided measures. Here all the base units are different, except the measure of time. This is the unit system you are most likely to encounter if SI units are not used. These units are used by the Americans and

British. As you can imagine, having different units in use from place to place makes scientific communication very difficult. This was the motivation for adopting a set of internationally agreed upon units.

Natural Units

This is the most sophisticated choice of units. Here the most fundamental discovered quantities (such as the speed of light) are set equal to 1. The argument for this choice is that all other quantities should be built from these fundamental units. This system of units is used in high energy physics and quantum mechanics.

1.5 The Importance of Units

Without units much of our work as scientists would be meaningless. We need to express our thoughts clearly and units give meaning to the numbers we calculate. Depending on which units we use, the numbers are different (e.g. 3.8 m and 3800 mm actually represent the same length). Units are an essential part of the language we use. Units must be specified when expressing physical quantities. In the case of the curtain example at the beginning of the chapter, the result of a misunderstanding would simply have been an incorrect amount of material cut. However, sometimes such misunderstandings have catastrophic results. Here is an extract from a story on CNN's website:

(NOTE TO SELF: This quote may need to be removed as the licence we are using allows for all parts of the document to be copied and I am not sure if this being copied is legit in all ways?)

> **NASA: Human error caused loss of Mars orbiter November 10, 1999**
>
> WASHINGTON (AP) — Failure to convert English measures to metric values caused the loss of the Mars Climate Orbiter, a spacecraft that smashed into the planet instead of reaching a safe orbit, a NASA investigation concluded Wednesday.
>
> The Mars Climate Orbiter, a key craft in the space agency's exploration of the red planet, vanished after a rocket firing September 23 that was supposed to put the spacecraft on orbit around Mars.
>
> An investigation board concluded that NASA engineers failed to convert English measures of rocket thrusts to newton, a metric system measuring rocket force. One English pound of force equals 4.45 newtons. A small difference between the two values caused the spacecraft to approach Mars at too low an altitude and the craft is thought to have smashed into the planet's atmosphere and was destroyed.
>
> The spacecraft was to be a key part of the exploration of the planet. From its station about the red planet, the Mars Climate Orbiter was to relay signals from the Mars Polar Lander, which is scheduled to touch down on Mars next month.
>
> "The root cause of the loss of the spacecraft was a failed translation of English units into metric units and a segment of ground-based, navigation-related mission software," said Arthus Stephenson, chairman of the investigation board.

This story illustrates the importance of being aware that different systems of units exist. Furthermore, we must be able to convert between systems of units!

1.6 Choice of Units

There are no wrong units to use, but a clever choice of units can make a problem look simpler. The vast range of problems makes it impossible to use a single set of units for everything without making some problems look much more complicated than they should. We can't easily compare the mass of the sun and the mass of an electron, for instance. This is why astrophysicists and atomic physicists use different systems of units.

We won't ask you to choose between different unit systems. For your present purposes the SI system is perfectly sufficient. In some cases you may come across quantities expressed in units other than the standard SI units. You will then need to convert these quantities into the correct SI units. This is explained in the next section.

1.7 How to Change Units— the "Multiply by 1" Technique

Firstly you obviously need some relationship between the two units that you wish to convert between. Let us demonstrate with a simple example. We will consider the case of converting millimetres (mm) to metres (m)— the SI unit of length. We know that there are $1000mm$ in $1m$ which we can write as

$$1000mm = 1m.$$

Now multiplying both sides by $\frac{1}{1000mm}$ we get

$$\frac{1}{1000mm}1000mm = \frac{1}{1000mm}1m,$$

which simply gives us

$$1 = \frac{1m}{1000mm}.$$

This is the conversion ratio from millimetres to metres. You can derive any conversion ratio in this way from a known relationship between two units. Let's use the conversion ratio we have just derived in an example:

Question: Express $3800mm$ in metres.

Answer:

$$\begin{aligned} 3800mm &= 3800mm \times 1 \\ &= 3800mm \times \frac{1m}{1000mm} \\ &= 3.8m \end{aligned}$$

Note that we wrote every unit in each step of the calculation. By writing them in and cancelling them properly, we can check that we have the right units when we are finished. We started with 'mm' and multiplied by '$\frac{m}{mm}$'. This cancelled the 'mm' leaving us with just 'm'— the SI unit we wanted to end up with! If we wished to do the reverse and convert metres to millimetres, then we would need a conversion ratio with millimetres on the top and metres on the bottom.

1.8 How Units Can Help You

We conclude each section of this book with a discussion of the units most relevant to that particular section. It is important to try to understand what the units mean. That is why thinking about the examples and explanations of the units is essential.

If we are careful with our units then the numbers we get in our calculations can be checked in a 'sanity test'.

1.8.1 What is a 'sanity test'?

This isn't a special or secret test. All we do is stop, take a deep breath, and look at our answer. Sure we always look at our answers— or do we? This time we mean stop and really look— does our answer make sense?

Imagine you were calculating the number of people in a classroom. If the answer you got was 1 000 000 people you would know it was wrong— that's just an insane number of people to have in a classroom. That's all a sanity check is— is your answer insane or not? But what units were we using? We were using people as our unit. This helped us to make sense of the answer. If we had used some other unit (or no unit) the number would have lacked meaning and a sanity test would have been much harder (or even impossible).

It is useful to have an idea of some numbers before we start. For example, let's consider masses. An average person has mass $70kg$, while the heaviest person in medical history had a mass of $635kg$. If you ever have to calculate a person's mass and you get $7000kg$, this should fail your sanity check— your answer is insane and you must have made a mistake somewhere. In the same way an answer of $0.00001kg$ should fail your sanity test.

The only problem with a sanity check is that you must know what typical values for things are. In the example of people in a classroom you need to know that there are usually 20–50 people in a classroom. Only then do you know that your answer of 1 000 000 must be wrong. Here is a table of typical values of various things (big and small, fast and slow, light and heavy— you get the idea):

Category	Quantity	Minimum	Maximum
People	Mass Height		

Table 1.4: Everyday examples to help with *sanity checks*

(NOTE TO SELF: Add to this table as we go along with examples from each section.)

Now you don't have to memorise this table but you should read it. The best thing to do is to refer to it every time you do a calculation.

1.9 Temperature

We need to make a special mention of the units used to describe temperature. The unit of temperature listed in Table 1.1 is not the everyday unit we see and use.

Normally the Celsius scale is used to describe temperature. As we all know, Celsius temperatures can be negative. This might suggest that any number is a valid temperature. In fact, the temperature of a gas is a measure of the average kinetic energy of the particles that make up the gas. As we lower the temperature so the motion of the particles is reduced until a point is reached where all motion ceases. The temperature at which this occurs is called *absolute zero*. There is no physically possible temperature colder than this. In Celsius, absolute zero is at -273^oC.

Physicists have defined a new temperature scale called the **Kelvin scale**. According to this scale absolute zero is at $0K$ and negative temperatures are not allowed. The size of one unit kelvin is exactly the same as that of one unit Celsius. This means that a change in temperature of 1 degree kelvin is equal to a change in temperature of 1 degree Celsius— the scales just start in different places. Think of two ladders with steps that are the same size but the bottom most step on the Celsius ladder is labelled -273, while the first step on the Kelvin ladder is labelled 0. There are still 100 steps between the points where water freezes and boils.

```
                         |----|   102 Celsius    |----|  375 Kelvin
                         |----|   101 Celsius    |----|  374 Kelvin
water boils   --->       |----|   100 Celsius    |----|  373 Kelvin
                         |----|   99  Celsius    |----|  372 Kelvin
                         |----|   98  Celsius    |----|  371 Kelvin
                                            .
                                            .
                                            .
                         |----|   2   Celsius    |----|  275 Kelvin
                         |----|   1   Celsius    |----|  274 Kelvin
ice melts     --->       |----|   0   Celsius    |----|  273 Kelvin
                         |----|   -1  Celsius    |----|  272 Kelvin
                         |----|   -2  Celsius    |----|  271 Kelvin
                                            .
                                            .
                                            .
                         |----|  -269 Celsius    |----|  4 Kelvin
                         |----|  -270 Celsius    |----|  3 Kelvin
                         |----|  -271 Celsius    |----|  2 Kelvin
                         |----|  -272 Celsius    |----|  1 Kelvin
absolute zero --->       |----|  -273 Celsius    |----|  0 Kelvin
```

(NOTE TO SELF: Come up with a decent picture of two ladders with the labels —water boiling and freezing—in the same place but with different labelling on the steps!)

This makes the conversion from kelvin to Celsius and back very easy. To convert from Celsius to kelvin add 273. To convert from kelvin to Celsius subtract 273. Representing the Kelvin temperature by T_K and the Celsius temperature by $T_{^\circ C}$,

$$T_K = T_{^\circ C} + 273. \tag{1.1}$$

It is because this conversion is additive that a difference in temperature of 1 degree Celsius is equal to a difference of 1 kelvin. The majority of conversions between units are multiplicative. For example, to convert from metres to millimetres we **multiply** by 1000. Therefore a change of $1m$ is equal to a change of $1000mm$.

1.10 Scientific Notation, Significant Figures and Rounding

(NOTE TO SELF: still to be written)

1.11 Conclusion

In this chapter we have discussed the importance of units. We have discovered that there are many different units to describe the same thing, although you should stick to SI units in your calculations. We have also discussed how to convert between different units. This is a skill you must acquire.

Chapter 2

Waves and Wavelike Motion

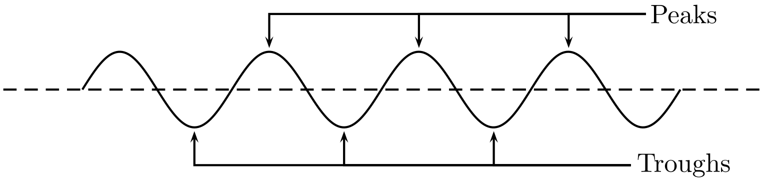
Waves occur frequently in nature. The most obvious examples are waves in water, on a dam, in the ocean, or in a bucket. We are most interested in the properties that waves have. All waves have the same properties so if we study waves in water then we can transfer our knowledge to predict how other examples of waves will behave.

2.1 What are waves?

Waves are disturbances which propagate (move) through a medium [1]. Waves can be viewed as a transfer energy rather than the movement of a particle. Particles form the medium through which waves propagate but they are not the wave. This will become clearer later.

Lets consider one case of waves: water waves. Waves in water consist of moving peaks and troughs. A peak is a place where the water rises higher than when the water is still and a trough is a place where the water sinks lower than when the water is still. A single peak or trough we call a *pulse*. A wave consists of a *train* of *pulses*.

So waves have peaks and troughs. This could be our first property for waves. The following diagram shows the peaks and troughs on a wave.



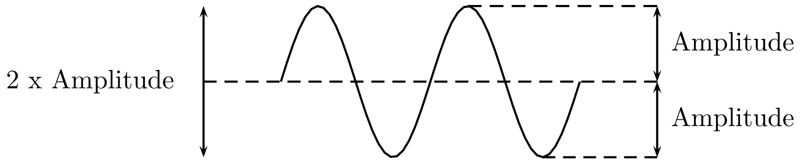
In physics we try to be as quantitative as possible. If we look very carefully we notice that the height of the peaks above the level of the still water is the same as the depth of the troughs below the level of the still water. The size of the peaks and troughs is the same.

2.1.1 Characteristics of Waves : Amplitude

The characteristic height of a peak and depth of a trough is called the *amplitude* of the wave. The vertical distance between the bottom of the trough and the top of the peak is twice the amplitude. We use symbols agreed upon by convention to label the characteristic quantities of

[1] Light is a special case, it exhibits wave-like properties but does not require a medium through which to propagate.

the waves. Normally the letter A is used for the amplitude of a wave. The units of amplitude are metres (m).



Worked Example 1

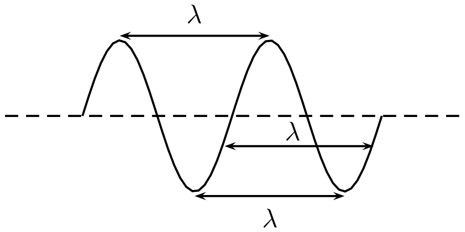
Question: (NOTE TO SELF: Make this a more exciting question) If the peak of a wave measures 2m above the still water mark in the harbour what is the amplitude of the wave?
Answer: The definition of the amplitude is the height that the water rises to above when it is still. This is exactly what we were told, so the answer is that the amplitude is 2m.

2.1.2 Characteristics of Waves : Wavelength

Look a little closer at the peaks and the troughs. The distance between two *adjacent* (next to each other) peaks is the same no matter which two adjacent peaks you choose. So there is a fixed distance between the peaks.

Looking closer you'll notice that the distance between two adjacent troughs is the same no matter which two troughs you look at. But, *more importantly*, its is the same as the distance between the peaks. This distance which is a characteristic of the wave is called the *wavelength.*

Waves have a characteristic wavelength. The symbol for the wavelength is λ. The units are metres (m).



The wavelength is the distance between any two adjacent points which are in *phase*. Two points in phase are separate by an integer (0,1,2,3,...) number of complete wave cycles. They don't have to be peaks or trough but they must be separated by a complete number of waves.

2.1.3 Characteristics of Waves : Period

Now imagine you are sitting next to a pond and you watch the waves going past you. First one peak, then a trough and then another peak. If you measure the time between two adjacent peaks you'll find that it is the same. Now if you measure the time between two adjacent troughs you'll find that its always the same, no matter which two adjacent troughs you pick. The time you have been measuring is the time for one wavelength to pass by. We call this time the period and it is a characteristic of the wave.

Waves have a characteristic time interval which we call the *period* of the wave and denote with the symbol T. It is the time it takes for any two adjacent points which are in phase to pass a fixed point. The units are seconds (s).

2.1.4 Characteristics of Waves : Frequency

There is another way of characterising the time interval of a wave. We timed how long it takes for one wavelength to pass a fixed point to get the period. We could also turn this around and say how many waves go by in 1 second.

We can easily determine this number, which we call the *frequency* and denote f. To determine the frequency, how many waves go by in 1s, we work out what fraction of a waves goes by in 1 second by dividing 1 second by the time it takes T. If a wave takes $\frac{1}{2}$ a second to go by then in 1 second two waves must go by. $\frac{1}{\frac{1}{2}} = 2$. The unit of frequency is the Hz or s^{-1}.

Waves have a characteristic frequency.

$$\boxed{f = \frac{1}{T}}$$

f : frequency (Hz or s^{-1})
T : period (s)

2.1.5 Characteristics of Waves : Speed

Now if you are watching a wave go by you will notice that they move at a constant velocity. The speed is the distance you travel divided by the time you take to travel that distance. This is excellent because we know that the waves travel a distance λ in a time T. This means that we can determine the speed.

$$\boxed{v = \frac{\lambda}{T}}$$

v : speed ($m.s^{-1}$)
λ : wavelength (m)
T : period (s)

There are a number of relationships involving the various characteristic quantities of waves. A simple example of how this would be useful is how to determine the velocity when you have the frequency and the wavelength. We can take the above equation and substitute the relationship between frequency and period to produce an equation for speed of the form

$$\boxed{v = f\lambda}$$

v : speed ($m.s^{-1}$)
λ : wavelength (m)
f : frequency (Hz or s^{-1})

Is this correct? Remember a simple first check is to check the units! On the right hand side we have velocity which has units ms^{-1}. On the left hand side we have frequency which is measured in s^{-1} multiplied by wavelength which is measure in m. On the left hand side we have ms^{-1} which is exactly what we want.

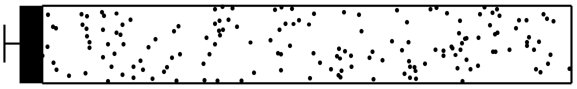
2.2 Two Types of Waves

We agreed that a wave was a moving set of peaks and troughs and we used water as an example. Moving peaks and troughs, with all the characteristics we described, in any medium constitute a wave. It is possible to have waves where the peaks and troughs are perpendicular to the direction of motion, like in the case of water waves. These waves are called *transverse waves.*

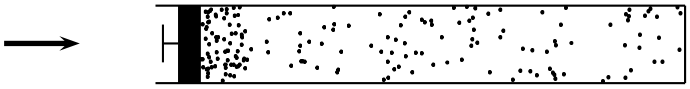
There is another type of wave. Called a *longitudinal wave* and it has the peaks and troughs in the same direction as the wave is moving. The question is how do we construct such a wave?

An example of a longitudinal wave is a pressure wave moving through a gas. The peaks in this wave are places where the pressure reaches a peak and the troughs are places where the pressure is a minimum.

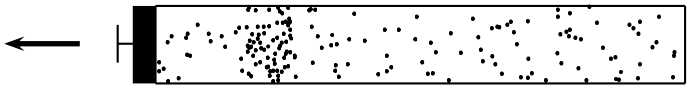
In the picture below we show the random placement of the gas molecules in a tube. The piston at the end moves into the tube with a repetitive motion. Before the first piston stroke the pressure is the same throughout the tube.



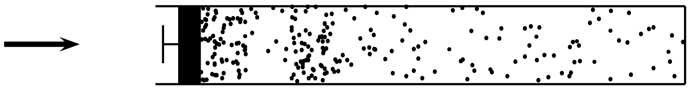
When the piston moves in it compresses the gas molecules together at the end of the tube. If the piston stopped moving the gas molecules would all bang into each other and the pressure would increase in the tube but if it moves out again fast enough then pressure waves can be set up.

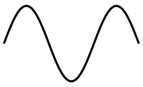


When the piston moves out again before the molecules have time to bang around then the increase in pressure moves down the tube like a pulse (single peak). The piston moves out so fast that a pressure trough is created behind the peak.



As this repeats we get waves of increased and decreased pressure moving down the tubes. We can describe these pulses of increased pressure (peaks in the pressure) and decreased pressure (troughs of pressure) by a sine or cosine graph.





There are a number of examples of each type of wave. Not all can be seen with the naked eye but all can be detected.

2.3 Properties of Waves

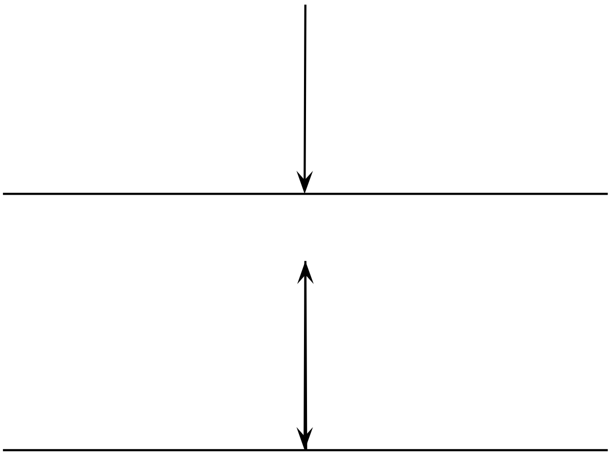
We have discussed some of the simple characteristics of waves that we need to know. Now we can progress onto some more interesting and, perhaps, less intuitive properties of waves.

2.3.1 Properties of Waves : Reflection

When waves strike a barrier they are *reflected*. This means that waves bounce off things. Sound waves bounce off walls, light waves bounce off mirrors, radar waves bounce off planes and it can explain how bats can fly at night and avoid things as small as telephone wires. The property of reflection is a very important and useful one.

(NOTE TO SELF: Get an essay by an air traffic controller on radar) (NOTE TO SELF: Get an essay by on sonar usage for fishing or for submarines)

When waves are reflected, the process of reflection has certain properties. If a wave hits an obstacle at a right angle to the surface (NOTE TO SELF: diagrams needed) then the wave is reflected directly backwards.

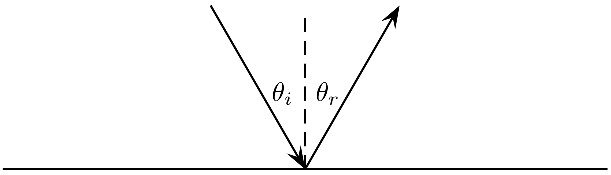


If the wave strikes the obstacle at some other angle then it is not reflected directly backwards. The angle that the waves arrives at is the same as the angle that the reflected waves leaves at. The angle that waves arrives at or is incident at equals the angle the waves leaves at or is reflected at. Angle of incidence equals angle of reflection

$$\theta_i = \theta_r \tag{2.1}$$

$$\theta_i = \theta_r$$

θ_i : angle of incidence
θ_r : angle of reflection



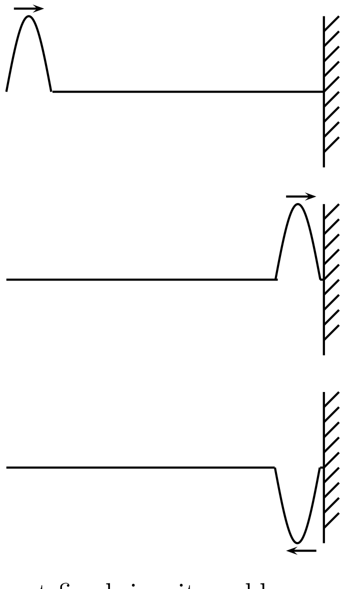
In the optics chapter you will learn that light is a wave. This means that all the properties we have just learnt apply to light as well. Its very easy to demonstrate reflection of light with a mirror. You can also easily show that angle of incidence equals angle of reflection.

If you look directly into a mirror your see yourself reflected directly back but if you tilt the mirror slightly you can experiment with different incident angles.

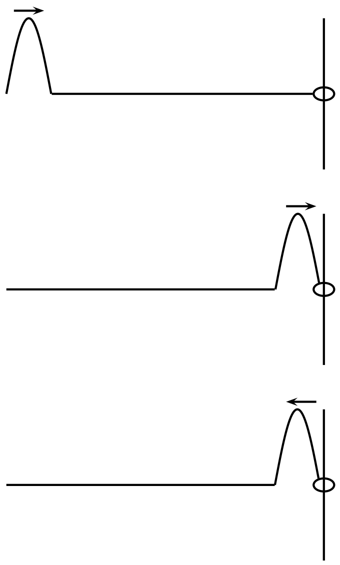
Phase shift of reflected wave

When a wave is reflected from a more dense medium it undergoes a phase shift. That means that the peaks and troughs are swapped around.

The easiest way to demonstrate this is to tie a piece of string to something. Stretch the string out flat and then flick the string once so a pulse moves down the string. When the pulse (a single peak in a wave) hits the barrier that the string is tide to it will be reflected. The reflected wave will look like a trough instead of a peak. This is because the pulse had undergone a phase change. The fixed end acts like an extremely dense medium.



If the end of the string was not fixed, i.e. it could move up and down then the wave would still be reflected but it would **not** undergo a phase shift. To draw a free end we draw it as a ring around a line. This signifies that the end is free to move.



2.3.2 Properties of Waves : Refraction

Sometimes waves move from one medium to another. The medium is the substance that is carrying the waves. In our first example this was the water. When the medium properties change it can affect the wave.

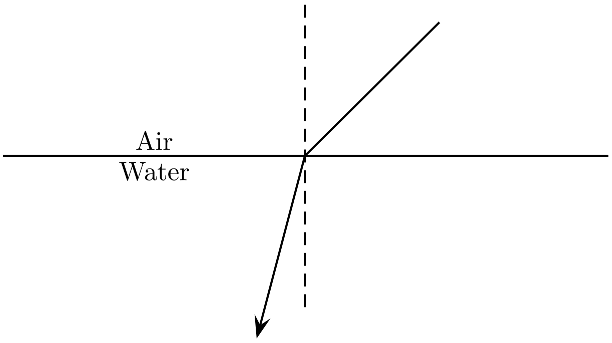
Let us start with the simple case of a water wave moving from one depth to another. The speed of the wave depends on the depth [2]. If the wave moves directly from the one medium to the other then we should look closely at the boundary. When a peak arrives at the boundary and moves across it must remain a peak on the other side of the boundary. This means that the peaks pass by at the same time intervals on either side of the boundary. The period and frequency remain the same! But we said the speed of the wave changes, which means that the distance it travels in one time interval is different i.e. the wavelength has **changed**.

Going from one medium to another the period or frequency does not change only the wavelength can change.

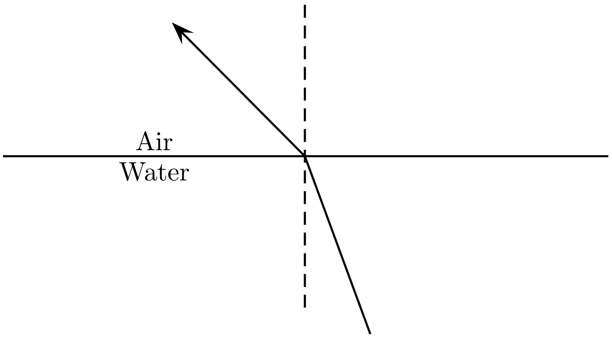
Now if we consider a water wave moving at an angle of incidence not 90 degrees towards a change in medium then we immediately know that not the whole wavefront will arrive at once. So if a part of the wave arrives and slows down while the rest is still moving faster before it arrives the angle of the wavefront is going to change. This is known as refraction. When a wave bends or changes its direction when it goes from one medium to the next.

If it slows down it turns towards the perpendicular.

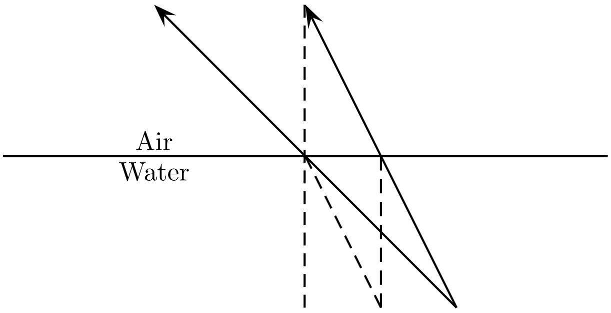
[2]



If the wave speeds up in the new medium it turns away from the perpendicular to the medium surface.



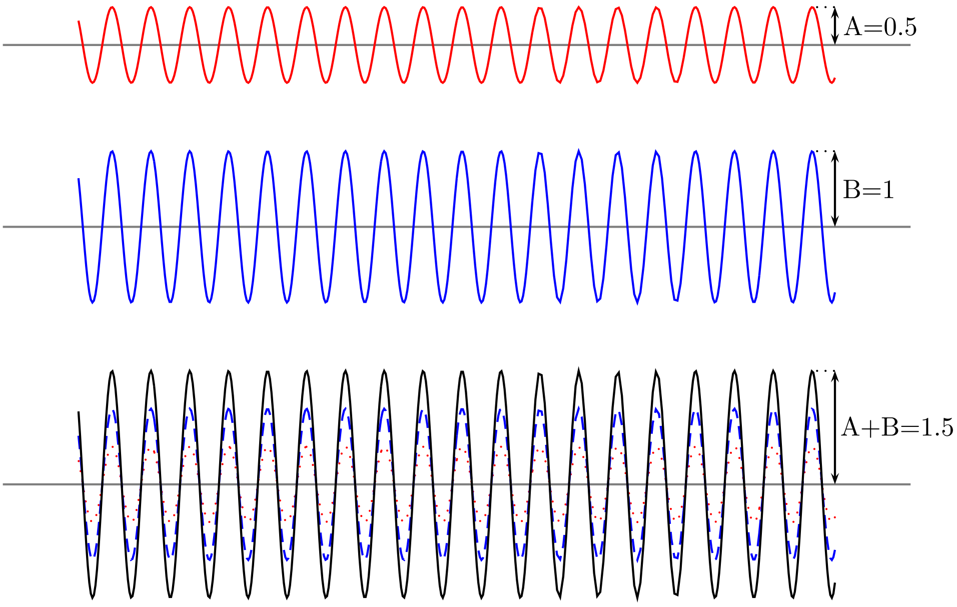
When you look at a stick that emerges from water it looks like it is bent. This is because the light from below the surface of the water bends when it leaves the water. Your eyes project the light back in a straight line and so the object looks like it is a different place.



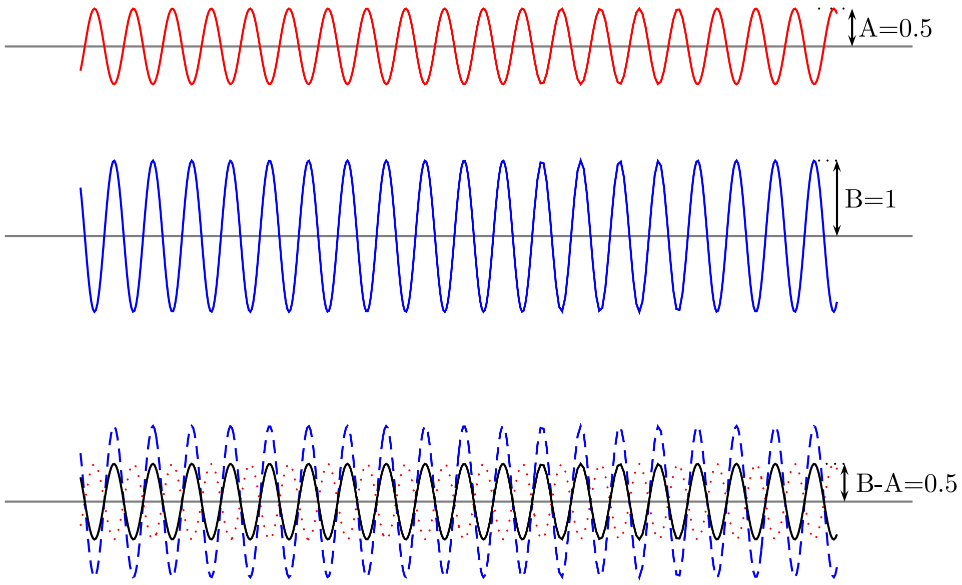
2.3.3 Properties of Waves : Interference

If two waves meet interesting things can happen. Waves are basically collective motion of particles. So when two waves meet they both try to impose their collective motion on the particles. This can have quite different results.

If two identical (same wavelength, amplitude and frequency) waves are both trying to form a peak then they are able to achieve the sum of their efforts. The resulting motion will be a peak which has a height which is the sum of the heights of the two waves. If two waves are both trying to form a trough in the same place then a deeper trough is formed, the depth of which is the sum of the depths of the two waves. Now in this case the two waves have been trying to do the same thing and so add together constructively. This is called *constructive interference*.



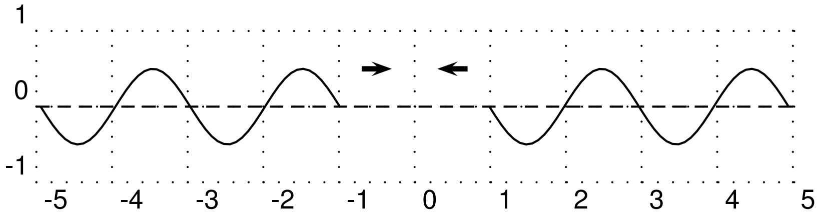
If one wave is trying to form a peak and the other is trying to form a trough then they are competing to do different things. In this case they can cancel out. The amplitude of the resulting wave will depend on the amplitudes of the two waves that are interfering. If the depth of the trough is the same as the height of the peak nothing will happen. If the height of the peak is bigger than the depth of the trough a smaller peak will appear and if the trough is deeper then a less deep trough will appear. This is *destructive interference*.



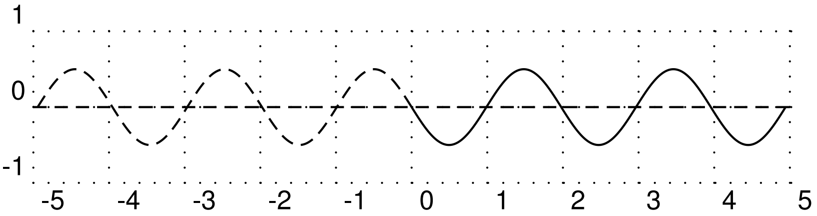
2.3.4 Properties of Waves : Standing Waves

When two waves move in opposite directions, through each other, interference takes place. If the two waves have the same frequency and wavelength then a specific type of constructive interference can occur: *standing waves* can form.

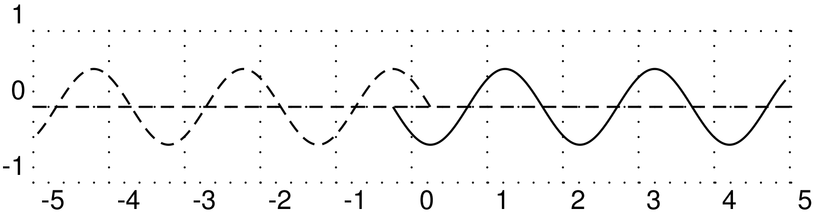
Standing waves are disturbances which don't appear to move, they look like they stay in the same place even though the waves that from them are moving. Lets demonstrate exactly how this comes about. Imagine a long string with waves being sent down it from either end. The waves from both ends have the same amplitude, wavelength and frequency as you can see in the picture below:



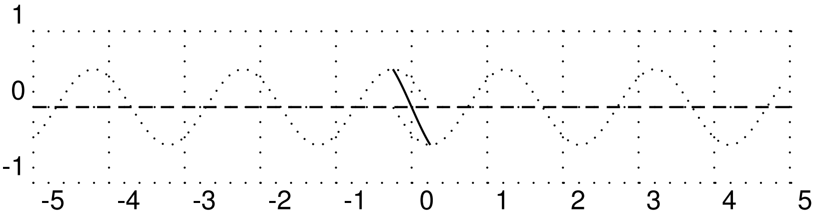
To stop from getting confused between the two waves we'll draw the wave from the left with a dashed line and the one from the right with a solid line. As the waves move closer together when they touch both waves have an amplitude of zero:



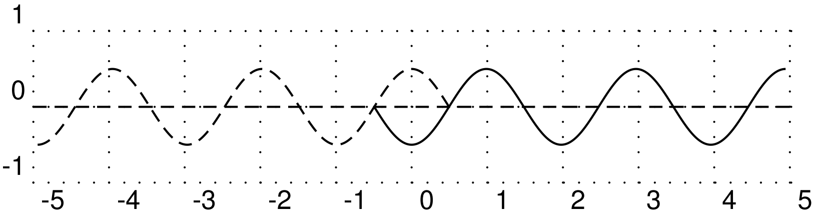
If we wait for a short time the ends of the two waves move past each other and the waves overlap. Now we know what happens when two waves overlap, we add them together to get the resulting wave.



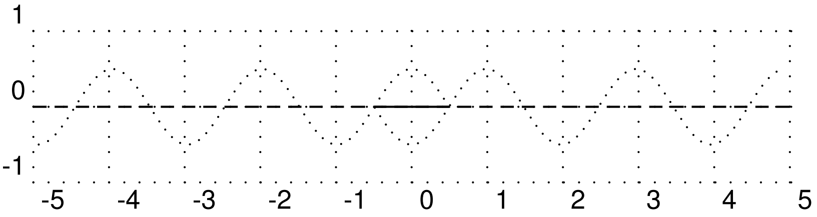
Now we know what happens when two waves overlap, we add them together to get the resulting wave. In this picture we show the two waves as dotted lines and the sum of the two in the overlap region is shown as a solid line:



The important thing to note in this case is that there are some points where the two waves always destructively interfere to zero. If we let the two waves move a little further we get the picture below:

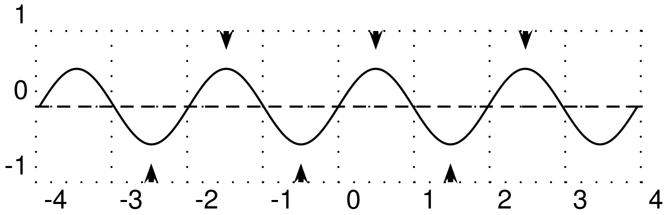


Again we have to add the two waves together in the overlap region to see what the sum of the waves looks like.

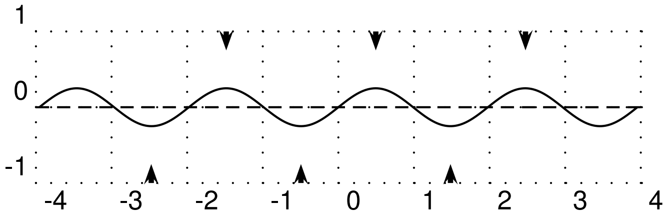


In this case the two waves have moved half a cycle past each other but because they are out of phase they cancel out completely. The point at 0 will always be zero as the two waves move past each other.

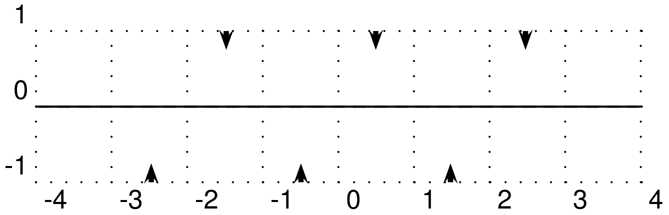
When the waves have moved past each other so that they are overlapping for a large region the situation looks like a wave oscillating in place. If we focus on the range -4, 4 once the waves have moved over the whole region. To make it clearer the arrows at the top of the picture show peaks where maximum positive constructive interference is taking place. The arrows at the bottom of the picture show places where maximum negative interference is taking place.



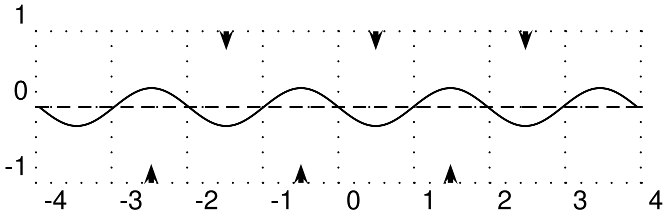
As time goes by the peaks become smaller and the troughs become shallower but they do not move.



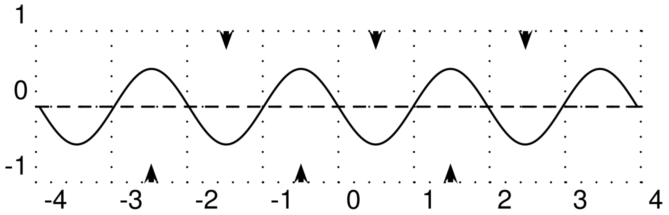
For an instant the entire region will look completely flat.



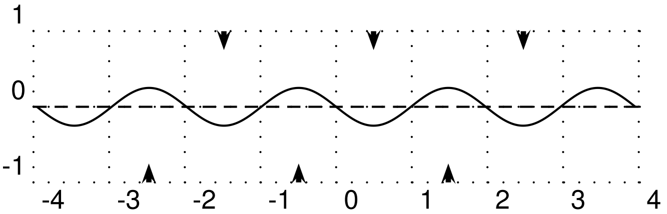
The various points continue their motion in the same manner.



Eventually the picture looks like the complete reflection through the x-axis of what we started with:

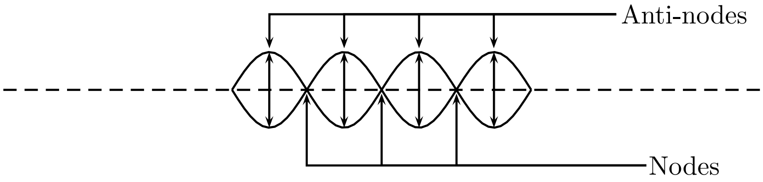


Then all the points begin to move back. Each point on the line is oscillating up and down with a different amplitude.

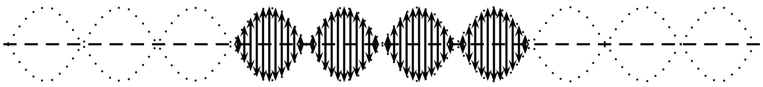


If we superimpose the two cases where the peaks were at a maximum and the case where the same waves were at a minimum we can see the lines that the points oscillate between. We call this the *envelope* of the standing wave as it contains all the oscillations of the individual points. A node is a place where the two waves cancel out completely as two waves destructively interfere in the same place. An anti-node is a place where the two waves constructively interfere.

Important: The distance between two anti-nodes is only $\frac{1}{2}\lambda$ because it is the distance from a peak to a trough in one of the waves forming the standing wave. It is the same as the distance between two adjacent nodes. This will be important when we workout the allowed wavelengths in tubes later. We can take this further because half-way between any two anti-nodes is a node. Then the distance from the node to the anti-node is half the distance between two anti-nodes. This is half of half a wavelength which is one quarter of a wavelength, $\frac{1}{4}\lambda$.



To make the concept of the envelope clearer let us draw arrows describing the motion of points along the line.



Every point in the medium containing a standing wave oscillates up and down and the amplitude of the oscillations depends on the location of the point. It is convenient to draw the envelope for the oscillations to describe the motion. We cannot draw the up and down arrows for every single point!

Reflection from a fixed end

If waves are reflected from a fixed end, for example tieing the end of a rope to a pole and then sending waves down it. The fixed end will always be a node. **Remember**: Waves reflected from a fixed end undergo a phase shift.

The wavelength, amplitude and speed of the wave cannot affect this, the fixed end is always a node.

Reflection from an open end

If waves are reflected from end, which is free to move, it is an anti-node. For example tieing the end of a rope to a ring, which can move up and down, around the pole. **Remember**: The waves sent down the string are reflected but do not suffer a phase shift.

Wavelengths of standing waves with fixed and open ends

There are many applications which make use of the properties of waves and the use of fixed and free ends. Most musical instruments rely on the basic picture that we have presented to create specific sounds, either through standing pressure waves or standing vibratory waves in strings.

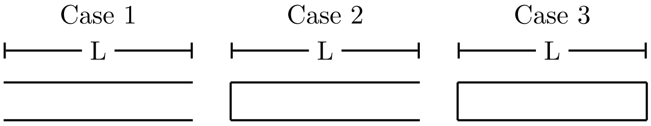
The key is to understand that a standing wave must be created in the medium that is oscillating. There are constraints as to what wavelengths can form standing waves in a medium.

For example, if we consider a tube of gas it can have

- both ends open (Case 1)
- one end open and one end closed (Case 2)
- both ends closed (Case 3).

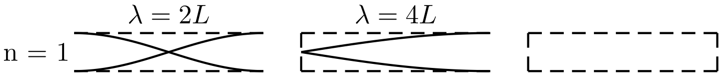
Each of these cases is slightly different because the open or closed end determines whether a node or anti-node will form when a standing wave is created in the tube. These are the primary constraints when we determine the wavelengths of potential standing waves. These constraints **must** be met.

In the diagram below you can see the three cases different cases. It is possible to create standing wave with different frequencies and wavelengths as long as the end criteria are met.



The longer the wavelength the less the number of anti-nodes in the standing waves. We cannot have a standing wave with 0 no anti-nodes because then there would be no oscillations. We use n to number to anti-nodes. If all of the tubes have a length L and we know the end constraints we can workout the wavelenth, λ, for a specific number of anti-nodes.

Lets workout the longest wavelength we can have in each tube, i.e. the case for n = 1.



Case 1: In the first tube both ends must be nodes so we can place one anti-node in the middle of the tube. We know the distance from one node to another is $\frac{1}{2}\lambda$ and we also know this distance is L. So we can equate the two and solve for the wavelength:

$$\begin{aligned} \frac{1}{2}\lambda &= L \\ \lambda &= 2L \end{aligned}$$

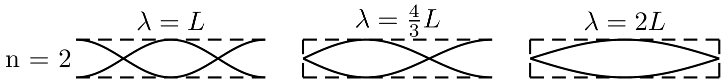
Case 2: In the second tube one ends must be a node and the other must be an anti-node. We are looking at the case with one anti-node we we are forced to have it at the end. We know the distance from one node to another is $\frac{1}{2}\lambda$ but we only have half this distance contained in the tube. So :

$$\begin{aligned} \frac{1}{2}(\frac{1}{2}\lambda) &= L \\ \lambda &= 4L \end{aligned}$$

Case 3: Here both ends are closed and so we must have two nodes so it is impossible to construct a case with only one node.

Next we determine which wavelengths could be formed if we had two nodes. Remember that we are dividing the tube up into smaller and smaller segments by having more nodes so we expect the wavelengths to get shorter.

NB: If you ever
a longer wave
for more nod
have made a m
Remember to
your answers ma



Case 1: Both ends are open and so they must be anti-nodes. We can have two nodes inside the tube only if we have one anti-node contained inside the tube and one on each end. This means we have 3 anti-nodes in the tube. The distance between any two anti-nodes is half a wavelength. This means there is half wavelength between the left side and the middle and another half wavelength between the middle and the right side so there must be one wavelength inside the tube. The safest thing to do is workout how many half wavelengths there are and equate this to the length of the tube L and then solve for λ.

Even though its very simple in this case we should practice our technique:

$$\begin{aligned} 2(\frac{1}{2}\lambda) &= L \\ \lambda &= L \end{aligned}$$

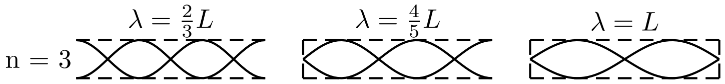
Case 2: We want to have two nodes inside the tube. The left end must be a node and the right end must be an anti-node. We can have one node inside the tube as drawn above. Again we can count the number of distances between adjacent nodes or anti-nodes. If we start from the left end we have one half wavelength between the end and the node inside the tube. The distance from the node inside the tube to the right end which is an anti-node is half of the distance to another node. So it is half of half a wavelength. Together these add up to the length of the tube:

$$\begin{aligned} \frac{1}{2}\lambda + \frac{1}{2}(\frac{1}{2}\lambda) &= L \\ \frac{2}{4}\lambda + \frac{1}{4}\lambda &= L \\ \frac{3}{4}\lambda &= L \\ \lambda &= \frac{4}{3}L \end{aligned}$$

Case 3: In this case both ends have to be nodes. This means that the length of the tube is one half wavelength: So we can equate the two and solve for the wavelength:

$$\begin{aligned} \frac{1}{2}\lambda &= L \\ \lambda &= 2L \end{aligned}$$

To see the complete pattern for all cases we need to check what the next step for case 3 is when we have an additional node. Below is the diagram for the case where n=3.



Case 1: Both ends are open and so they must be anti-nodes. We can have three nodes inside the tube only if we have two anti-node contained inside the tube and one on each end. This means we have 4 anti-nodes in the tube. The distance between any two anti-nodes is half a wavelength. This means there is half wavelength between every adjacent pair of anti-nodes. We count how many gaps there are between adjacent anti-nodes to determine how many half wavelengths there are and equate this to the length of the tube L and then solve for λ.

$$\begin{aligned} 3(\frac{1}{2}\lambda) &= L \\ \lambda &= \frac{2}{3}L \end{aligned}$$

Case 2: We want to have three nodes inside the tube. The left end must be a node and the right end must be an anti-node, so there will be two nodes between the ends of the tube. Again we can count the number of distances between adjacent nodes or anti-nodes, together these add up to the length of the tube. Remember that the distance between the node and an adjacent anti-node is only half the distance between adjacent nodes. So starting from the left end we 3 nodes so 2 half wavelength intervals and then only a node to anti-node distance:

$$\begin{aligned} 2(\frac{1}{2}\lambda) + \frac{1}{2}(\frac{1}{2}\lambda) &= L \\ \lambda + \frac{1}{4}\lambda &= L \\ \frac{5}{4}\lambda &= L \\ \lambda &= \frac{4}{5}L \end{aligned}$$

Case 3: In this case both ends have to be nodes. With one node in between there are two sets of adjacent nodes. This means that the length of the tube consists of two half wavelength sections:

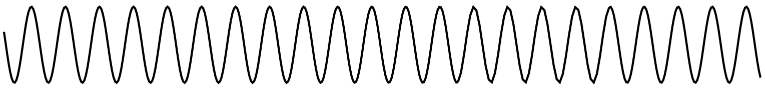
$$\begin{aligned} 2(\frac{1}{2}\lambda) &= L \\ \lambda &= L \end{aligned}$$

2.3.5 Beats

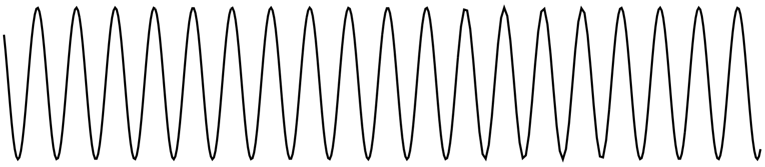
If the waves that are interfering are not identical then the waves form a modulated pattern with a changing amplitude. The peaks in amplitude are called beats. If you consider two sound waves interfering then you hear sudden beats in loudness or intensity of the sound.

The simplest illustration is two draw two different waves and then add them together. You can do this mathematically and draw them yourself to see the pattern that occurs.

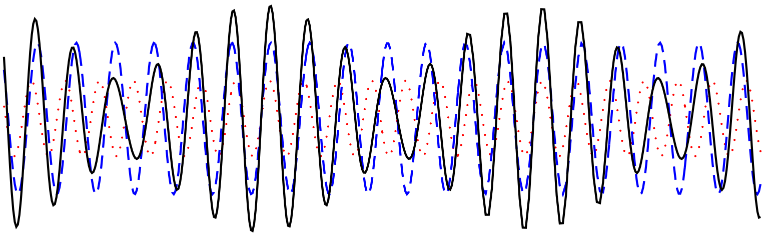
Here is wave 1:



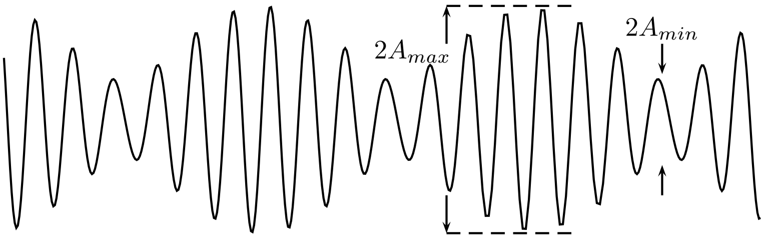
Now we add this to another wave, wave 2:



When the two waves are added (drawn in coloured dashed lines) you can see the resulting wave pattern:



To make things clearer the resulting wave without the dashed lines is drawn below. Notice that the peaks are the same distance apart but the amplitude changes. If you look at the peaks they are modulated i.e. the peak amplitudes seem to oscillate with another wave pattern. This is what we mean by modulation.



The maximum amplitude that the new wave gets to is the sum of the two waves just like for constructive interference. Where the waves reach a maximum it is constructive interference.

The smallest amplitude is just the difference between the amplitudes of the two waves, exactly like in destructive interference.

The beats have a frequency which is the difference between the frequency of the two waves that were added. This means that the beat frequency is given by

$$f_B = |f_1 - f_2| \tag{2.2}$$

$$\boxed{f_B = |f_1 - f_2|}$$

f_B	: beat frequency (Hz or s^{-1})
f_1	: frequency of wave 1 (Hz or s^{-1})
f_2	: frequency of wave 2 (Hz or s^{-1})

2.3.6 Properties of Waves : Diffraction

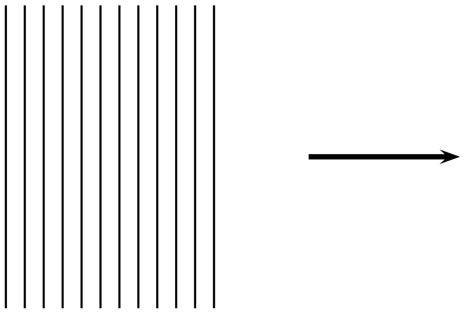
One of the most interesting, and also very useful, properties of waves is *diffraction*. When a wave strikes a barrier with a hole only part of the wave can move through the hole. If the hole is similar in size to the wavelength of the wave diffractions occurs. The waves that comes through the hole no longer looks like a straight wave front. It bends around the edges of the hole. If the hole is small enough it acts like a point source of circular waves.

This bending around the edges of the hole is called diffraction. To illustrate this behaviour we start by with Huygen's principle.

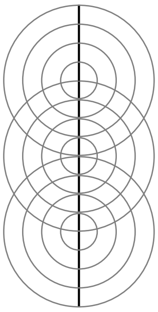
Huygen's Principle

Huygen's principle states that each point on a wavefront acts like a point source or circular waves. The waves emitted from each point interfere to form another wavefront on which each point forms a point source. A long straight line of points emitting waves of the same frequency leads to a straight wave front moving away.

To understand what this means lets think about a whole lot of peaks moving in the same direction. Each line represents a peak of a wave.

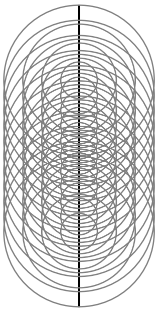


If we choose three points on the next wave front in the direction of motion and make each of them emit waves isotropically (i.e. the same in all directions) we will get the sketch below:

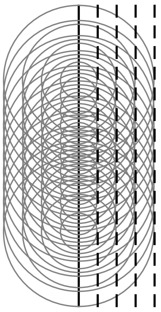


What we have drawn is the situation if those three points on the wave front were to emit waves of the same frequency as the moving wave fronts. Huygens principle says that every point on the wave front emits waves isotropically and that these waves interfere to form the next wave front.

To see if this is possible we make more points emit waves isotropically to get the sketch below:



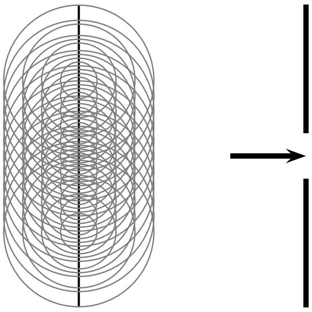
You can see that the lines from the circles (the peaks) start to overlap in straight lines. To make this clear we redraw the sketch with dashed lines showing the wavefronts which would form. Our wavefronts are not perfectly straight lines because we didn't draw circles from every point. If we had it would be hard to see clearly what is going on.



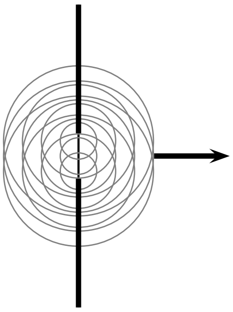
Huygen's principle is a method of analysis applied to problems of wave propagation. It recognizes that each point of an advancing wave front is in fact the center of a fresh disturbance and the source of a new train of waves; and that the advancing wave as a whole may be regarded as the sum of all the secondary waves arising from points in the medium already traversed. This view of wave propagation helps better understand a variety of wave phenomena, such as diffraction.

Wavefronts Moving Through an Opening

Now if allow the wavefront to impinge on a barrier with a hole in it, then only the points on the wavefront that move into the hole can continue emitting forward moving waves - but because a lot of the wavefront have been removed the points on the edges of the hole emit waves that bend round the edges.



The wave front that impinges (strikes) the wall cannot continue moving forward. Only the points moving into the gap can. If you employ Huygens' principle you can see the effect is that the wavefronts are no longer straight lines.



For example, if two rooms are connected by an open doorway and a sound is produced in a remote corner of one of them, a person in the other room will hear the sound as if it originated at the doorway. As far as the second room is concerned, the vibrating air in the doorway is the source of the sound. The same is true of light passing the edge of an obstacle, but this is not as easily observed because of the short wavelength of visible light.

This means that when waves move through small holes they appear to bend around the sides because there aren't enough points on the wavefront to form another straight wavefront. This is bending round the sides we call *diffraction*.

2.3.7 Properties of Waves : Dispersion

Dispersion is a property of waves where the speed of the wave through a medium depends on the frequency. So if two waves enter the same dispersive medium and have different frequencies they will have different speeds in that medium even if they both entered with the same speed.

We will come back to this topic in optics.

2.4 Practical Applications of Waves: Sound Waves

2.4.1 Doppler Shift

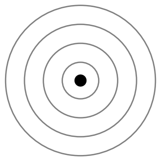
The Doppler shift is an effect which becomes apparent when the source of sound waves or the person hearing the sound waves is moving. In this case the frequency of the sounds waves can be different.

This might seem strange but you have probably experienced the doppler effect in every day life. When would you notice it. The effect depends on whether the source of the sound is moving away from the listener or if it is moving towards the listener. If you stand at the side fo the road or train tracks then a car or train driving by will at first be moving towards you and then away. This would mean that we would experience the biggest change in the effect.

We said that it effects the frequency of the sound so the sounds from the car or train would sound different, have a different frequency, when the car is coming towards you and when it is moving away from you.

Why does the frequency of the sound change when the car is moving towards or away from you? Lets convince ourselves that it must change!

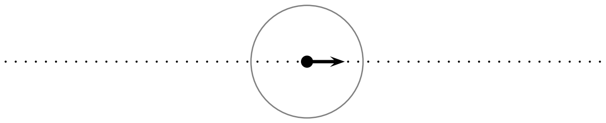
Imagine a source of sound waves with constant frequency and amplitude. Just like each of the points on the wave front from the Huygen's principle section.



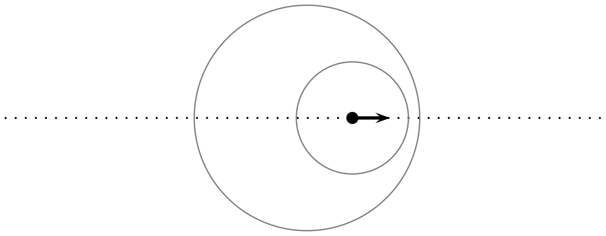
Remember the sound waves are disturbances moving through the medium so if the source moves or stops after the sound has been emitted it can't affect the waves that have been emitted already.

The Doppler shift happens when the source moves while emitting waves. So lets imagine we have the same source as above but now its moving to the right. It is emitting sound at a constant frequency and so the time between peaks of the sound waves will be constant but the position will have moved to the right.

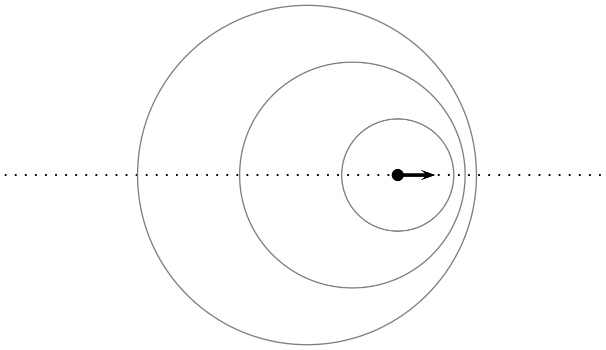
In the picture below we see that our sound source (the black circle) has emitted a peak which moves away at the same speed in all directions. The source is moving to the right so it catches up a little bit with the peak that is moving away to the right.



When the second peak is emitted it is from a point that has moved to the right. This means that the new or second peak is closer to the first peak on the right but further away from the first peak on the left.



If the source continues moving at the same speed in the same direction (i.e. with the same velocity which you will learn more about later). then the distance between peaks on the right of the source is the constant. The distance between peaks on the left is also constant but they are different on the left and right.



This means that the time between peaks on the right is less so the frequency is higher. It is higher than on the left and higher than if the source were not moving at all.

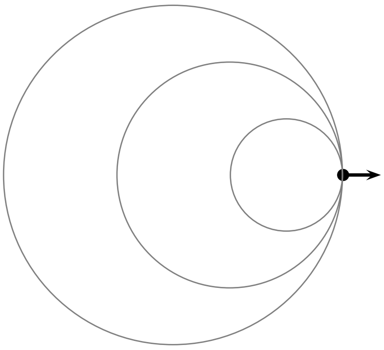
On the left hand side the peaks are further apart than on the right and further apart than if the source were at rest - this means the frequency is lower.

So what happens when a car drives by on the freeway is that when it approaches you hear higher frequency sounds and then when it goes past you it is moving away so you hear a lower frequency.

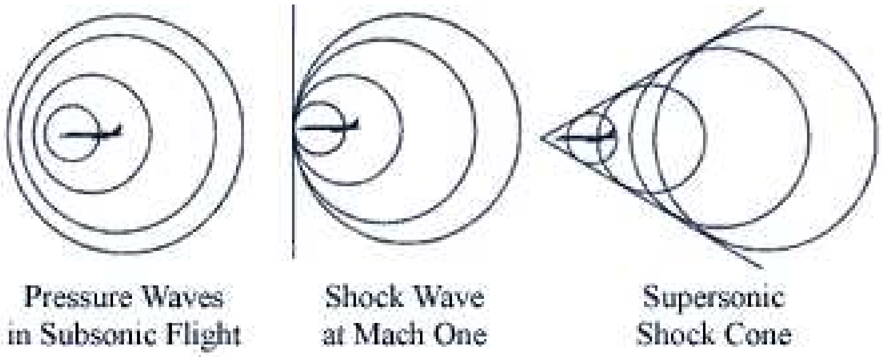
2.4.2 Mach Cone

Now we know that the waves move away from the source at the speed of sound. What happens if the source moves at the speed of sound?

This means that the wave peaks on the right never get away from the source so each wave is emitted on top of the previous on on the right hand side like in the picture below.



If the source moves faster than the speed of sound a cone of wave fronts is created. This is called a Mach cone. Sometimes we use the speed of sound as a reference to describe the speed of the object. So if the object moves at exactly the speed of sound we can say that it moves at 1 times the speed of sound. If it moves at double the speed of sound then we can say that it moves at 2 times the speed of sound. A convention for this is to use Mach numbers, so moving at the speed of sound is Mach one (one times the speed of sound) and moving at twice the speed of sound is called Mach two (twice the speed of sound).



2.4.3 Ultra-Sound

Ultrasound is sound with a frequency greater than the upper limit of human hearing, approximately 20 kilohertz. Some animals, such as dogs, dolphins, and bats, have an upper limit that is greater than that of the human ear and can hear ultrasound.

Ultrasound has industrial and medical applications. Medical ultrasonography can visualise muscle and soft tissue, making them useful for scanning the organs, and obstetric ultrasonography is commonly used during pregnancy. Typical diagnostic ultrasound scanners operate in the frequency range of 2 to 13 megahertz. More powerful ultrasound sources may be used to generate local heating in biological tissue, with applications in physical therapy and cancer treatment. Focused ultrasound sources may be used to break up kidney stones.

Ultrasonic cleaners, sometimes called supersonic cleaners, are used at frequencies from 20-40 kHz for jewellery, lenses and other optical parts, watches, dental instruments, surgical instruments and industrial parts.

These cleaners consist of containers with a fluid in which the object to be cleaned is placed. Ultrasonic waves are then sent into the fluid. The main mechanism for cleaning action in an ultrasonic cleaner is actually the energy released from the collapse of millions of microscopic cavitation events occurring in the liquid of the cleaner.

Interesting Fact: Ultrasound generator/speaker systems are sold with claims that they frighten away rodents and insects, but there is no scientific evidence that the devices work; controlled tests have shown that rodents quickly learn that the speakers are harmless.

Medical Ultrasonography

Medical ultrasonography makes uses the fact that waves are partially reflected when the medium in which they are moving changes density.

If the density increases then the reflected waves undergoes a phase shift exactly like the case where the waves in a string were reflected from a fixed end. If the density decreases then the reflected waves has the same phase exactly like the case where the waves in a string were reflected from a free end.

Combining these properties of waves with modern computing technology has allowed medical professionals to develop an imaging technology to help with many aspects of diagnosis. Typical ultrasound units have a hand-held probe (often called a scan head) that is placed directly on and moved over the patient: a water-based gel ensures good contact between the patient and scan head.

Ultrasonic waves are emitted from the scan head and sent into the body of the patient. The scan head also acts a receiver for reflected waves. From detailed knowledge of interference and reflection an image of the internal organs can be constructed on a screen by a computer programmed to process the reflected signals.

Interesting Fact: Medical ultrasonography was invented in 1953 at Lund University by cardiologist Inge Edler and Carl Hellmuth Hertz, the son of Gustav Ludwig Hertz, who was a graduate student at the department for nuclear physics.

Uses

Ultrasonography is widely utilized in medicine, primarily in gastroenterology, cardiology, gynaecology and obstetrics, urology and endocrinology. It is possible to perform diagnosis or therapeutic procedures with the guidance of ultrasonography (for instance biopsies or drainage of fluid collections).

Strengths of ultrasound imaging

- It images muscle and soft tissue very well and is particularly useful for finding the interfaces between solid and fluid-filled spaces.
- It renders ”live” images, where the operator can dynamically select the most useful section for diagnosing and documenting changes, often enabling rapid diagnoses.
- It shows the structure as well as some aspects of the function of organs.

- It has no known long-term side effects and rarely causes any discomfort to the patient.
- Equipment is widely available and comparatively flexible; examinations can be performed at the bedside.
- Small, easily carried scanners are available.
- Much cheaper than many other medical imaging technology.

Weaknesses of ultrasound imaging

- Ultrasound has trouble penetrating bone and performs very poorly when there is air between the scan head and the organ of interest. For example, overlying gas in the gastrointestinal tract often makes ultrasound scanning of the pancreas difficult.
- Even in the absence of bone or air, the depth penetration of ultrasound is limited, making it difficult to image structures that are far removed from the body surface, especially in obese patients.
- The method is operator-dependent. A high level of skill and experience is needed to acquire good-quality images and make accurate diagnoses.

Doppler ultrasonography

Ultrasonography can be enhanced with Doppler measurements, which employ the Doppler effect to assess whether structures (usually blood) are moving towards or away from the probe. By calculating the frequency shift of a particular sample volume, e.g. within a jet of blood flow over a heart valve, its speed and direction can be determined and visualised. This is particularily useful in cardiovascular studies (ultrasonography of the vasculature and heart) and essential in many areas such as determining reverse blood flow in the liver vasculature in portal hypertension. The Doppler information is displayed graphically using spectral Doppler, or as an image using colour Doppler or power Doppler. It is often presented audibly using stereo speakers: this produces a very distinctive, despite synthetic, sound.

2.5 Important Equations and Quantities

Frequency:

$$f = \frac{1}{T}. \tag{2.3}$$

Quantity	Symbol	S.I. Units	Direction
Amplitude	A	m	—
Period	T	s	—
Wavelength	λ	m	—
Frequency	f	Hz **or** s^{-1}	—
Speed	v	$m.s^{-1}$	—

Table 2.1: Units used in **Waves**

Speed:

$$\begin{aligned} v &= f\lambda \\ &= \frac{\lambda}{T} \end{aligned}$$

Chapter 3

Geometrical Optics

In this chapter we will study geometrical optics. Optics is the branch of physics that describes the behavior and properties of light and the interaction of light with matter. By geometrical optics we mean that we will study all the optics that we can treat using geometrical analysis (geometry), i.e. lines, angles and circles.

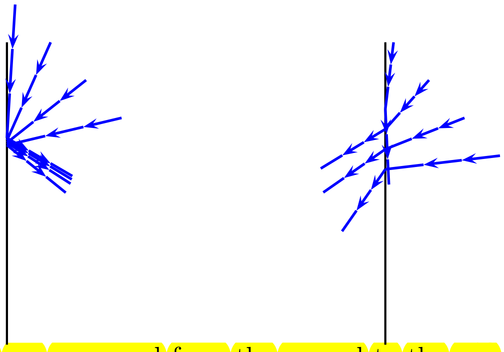
As we have seen in the previous chapters, light propagates as a wave. The waves travel in straight lines from the source. So we will consider light as a set of rays. The wavelike nature will become apparent when the waves go from one medium to another. Using rays and Snell's law to describe what happens when the light ray moves from one medium to another we can solve all the geometrical optics problems in this chapter.

3.1 Refraction re-looked

We have seen that waves refract as they move from shallower to deeper water or vise versa, thus light also refracts as it moves between two mediums of different densities.

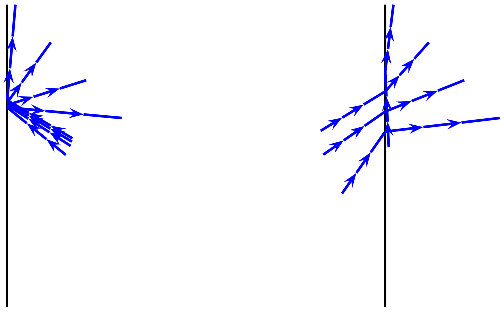
We may consider the parallel beams of light in Fig 1(a) as a set of wheels connected by a straight rod placed through their centres as in Fig1(b). As the wheels roll onto the grass (representing higher density), they begin to slow down, while those still on the tarmac move at a relatively faster speed. This shifts the direction of movtion towards the normal of the grass-tarmac barrier. Likewise light moving from a medium of low density to a medium of high density (Fig 2) moves towards the normal, hence the angle of incidence (i) is greater than the angle of Refraction (r): i ¿ r for d1 ¡ d2

(NOTE TO SELF: Diagram of ray moving into more dense medium) (NOTE TO SELF: Discussion of Snell's Law) (NOTE TO SELF: Follow by discussion of how we can reverse rays) (NOTE TO SELF: Show its all the same)



where both angles are measured from the normal to the ray. I is the incident ray and R is the refracted ray.

Rays are reversible

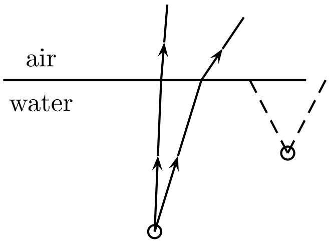


3.1.1 Apparent and Real Depth:

If you submerge a straight stick in water and look at the stick where it passes through the surface of the water you will notice that it looks like the stick bends. If you remove you will see that the stick did not bend. The bending of the stick is a very common example of refraction.

How can we explain this? We can start with a simple object under water. We can see things because light travels from the object into our eyes where it is absorbed and sent to the brain for processing. The human brain interprets the information it receives using the fact that light travels in straight lines. This is why the stick looks bent. The light did not travel in a straight line from the stick underwater to your eye. It was refracted at the surface. Your brain assumes the light travelled in a straight line and so it intrepets the information so that it thinks the stick is in a different place.

This phenomenon is easily explained using ray diagrams as in Fig**??**. The real light rays are represented with a solid line while dashed lines depict the virtual rays. The real light rays undergo refraction at the surface of the water hence move away from the normal. However the eye assumes that light rays travel in straight lines, thus we extend the refracted rays until they converge to a point. These are virtual rays as in reality the light was refracted and did not originate from that point.



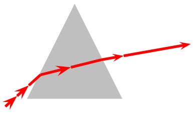
(NOTE TO SELF: Insert an example with water - can be worked example!)

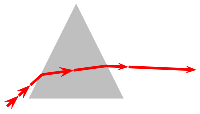
We note that the image of is seen slightly higher and ahead of the object. Where would we see the object if it was submerged in a fluid denser than water?

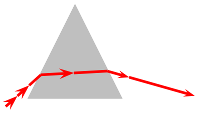
(NOTE TO SELF: Insert an example with denser medium show change - can be worked example!)

3.1.2 Splitting of White Light

How is a rainbow formed? White is a combination of all other colours of light. Each colour has a different wavelength and is thus diffracted through different angles.

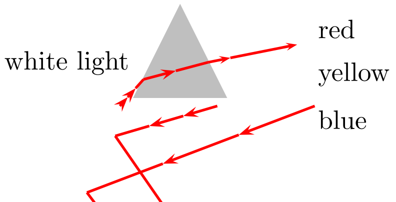

red


yellow


blue

The splitting of white light into its component colours may be demonstrated using a triangular prism (Fig4). White light is incident on the prism. As the white light enters the glass its component colours are diffracted through different angles. This separation is further expanded as the light rays leave the prism. Why? What colour is diffracted the most? If red has the longest wavelength and violet the shortest, what is the relation between refraction and wavelength?

As the sun appears after a rainstorm, droplets of water still remain in the atmosphere. These act as tiny prisms that split the suns light up into its component colours forming a rainbow.



3.1.3 Total Internal Reflection

Another useful application of refraction is the periscope. We know that as light move from higher to lower density mediums, light rays tend to be diffracted towards the normal. We also know that the angle of refraction is greater than the angle of incidence and increases as we increase the angle of incidence (Fig 5(a),(b)). At a certain angle of incidence, c, the refracted angle equals 90o (Fig 5(c)), this angle is called the critical angle. For any angle of incidence greater than c the light will be refracted back into the incident medium, such refraction is called Total Internal reflection (Fig 5(d)).

A periscope uses two 90o triangular prisms as total internal reflectors. Light enters the periscope and is reflected by the first prism down the chamber, where again the light is reflected to the observer. This may be illustrated using a ray diagram as in (Fig 6).

3.2 Lenses

Lenses are used in many aspects of technology ranging from contact lenses to projectors.

We shall again use light rays to explain the properties on lenses. There are two types of lenses, namely: Converging: Lenses that cause parallel light rays to converge to a point (Fig 7). Diverging: Lenses that cause parallel light rays to diverge (Fig 8). Such deviation of light rays are caused by refraction.

It is now a good time to introduce a few new definitions: Optical Centre: The centre of the lens Principal Axis: The line joining two centres of curvature of the surfaces of the lens. Focal Point: The point at which light rays, parallel to the principal axis, converge, or appear to converge, after passing through the lens. Focal Length: The distance between the optical centre and the focal point.

As seen in Fig 7 and 8, the focal point of a converging lens is real, while the focal point of a diverging lens is vitual.

3.2.1 Convex lenses

Convex lenses are in general converging lenses. Hence they possess real focal points (with one exception, discussed later). Such focal points allow for real images to be formed.

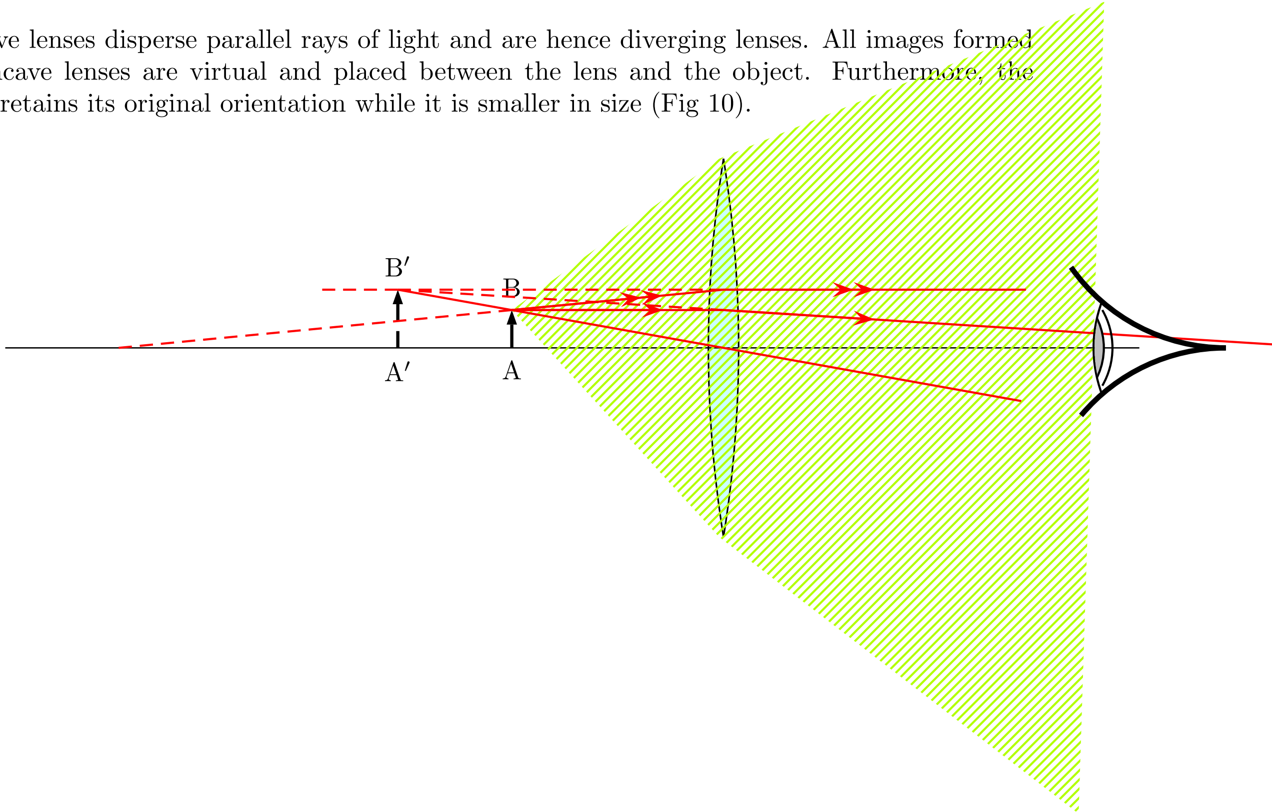
One may verify the above by placing an illuminated object on one side of a convex lens, ensuring that the distance between them is greater than the focal length (explained later). By placing a screen on the other side of the lens and varing the distance, one will acquire a sharp image of the object on the screen.

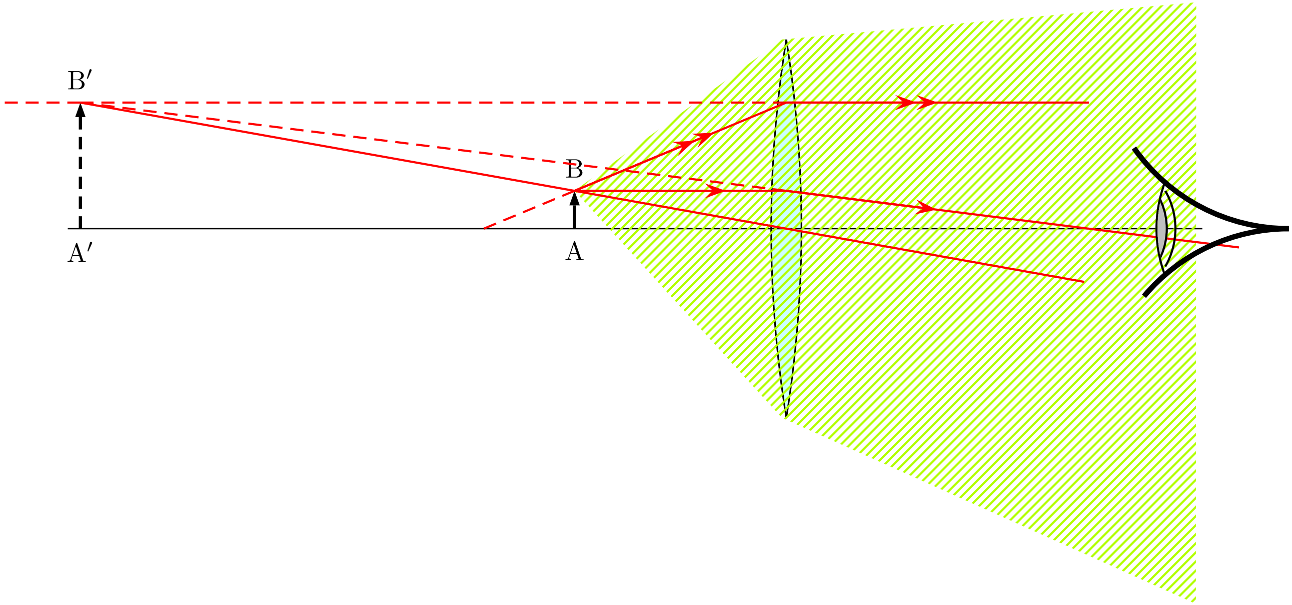
The ray diagrams below illustrate the images formed when an object is placed a distance d from the optical centre of a convex lens with focal length F.

d Image type Magnification Orientation Image Position (I) Figure ¿2F Real ¡1 Inverted F¡I¡2F 9(a) 2F Real 1 Inverted 2F 9(b) F¡d¡2F Real ¿1 Inverted ¿2F 9(c) F No Image - - - 9(d) ¡F Virtual ¿1 Original ¿d 9(e)

3.2.2 Concave Lenses

Concave lenses disperse parallel rays of light and are hence diverging lenses. All images formed by concave lenses are virtual and placed between the lens and the object. Furthermore, the image retains its original orientation while it is smaller in size (Fig 10).





3.2.3 Magnification

By Definition the magnification, m, is:

m = (Height of Image) / (Height of Object)

however, using similar triangles, we may prove that:

m = (Distance of Image) / (Distance of Object)

where both distances are measured from the optical centre.

The above method allows us to accurately estimate the magnification of an image. This is commonly used in a compound microscope as discussed in the next section.

Worked Example 2

A

n object is placed 0.5m in front of a convex a lens.

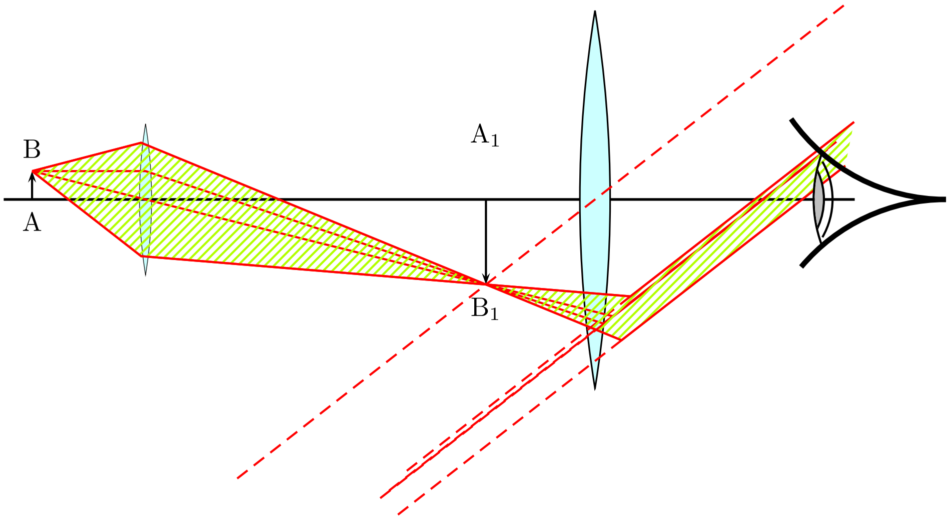
- At what distance should a screen be placed in order to create an image with a magnification of 1.5?
- If the height of the image and object are equal, what is the focal length of the lens?

Solutions: a)m = (Distance of Image) / (Distance of Object)
therefore Distance of Image = (Distance of Object) x m
Distance of Image = 0.5m x 1.5 = 0.75m
b)This implies that the magnification m = 1. Therefore from the above table it is seen that d = I = 2F Therefore F = d/2 = 0.25m

3.2.4 Compound Microscope

This type of microscope uses two convex lenses. The first creates a real magnified image of the object that is in turn used by the second lens to create the final image. This image is virtual and again enlarged. The final image, as seen in Fig 11, is virtual, enlarged and inverted.

The lens, L1 that forms the real image is called the objective lens, while L2 is referred to as the eyepiece.



3.2.5 The Human Eye

The eye also contains a biconvex lens that is used to focus objects onto the retina of the eye. However, in some cases the lens maybe abnormal and cause defects in ones vision.

Hyperopia (Long-Sightedness)

This occurs when the image is focussed beyond the retina. Hyperopia is due to the eyeball being too short or the lens not being convex enough. A convex lens is used to correct this defect (Fig12).

Miopia (Short-Sightedness)

Images are focussed before the retina. The lens being too convex or the eyeball being too long causes this. A concave lens corrects this defect (Fig 13).

Astigmatism

When this occurs, one is able to focus in one plane (e.g. vertical) but not another. This again is due to a distortion in the lens and may be cured by using relevant lenses.

3.3 Introduction

Light is at first, something we feel incredibly familiar with. It can make us feel warm, it allows us to see, allows mirrors and lenses to 'work', allows for ...

Under more careful study light exhibits many fascinating and wonderful properties. The study of light has led to many important and amazing scientific discoveries and much progress. For example fibre optics, lasers and CCD's play a huge role in modern technology.

- Light is a form of energy. This is demonstrated by Crookes Radiometer or Solar Cells
- Light travels in straight lines. This is demonstrated by an experiment involving a card with an hole looking at light source e.g. candle. Also the simple camera: candle black box one side pin hole other side grease proof paper. We see the candle upside down on the paper
- Light travels at a constant speed. The speed depends on the medium it is in. (substance like air or water)[1].
- Nothing travels faster than the speed of light in a vacuum $c = 2.99790 \times 10^8$. This is one of the fundamental constants in physics.

Probably the most important use of light by the majority of living things on earth is that it allows them to see. If the light from an object enters our optical detectors/sensors *i.e.* **our eyes!**, we can see that object. In some cases the light originates at the object itself. Objects which give out light are said to be luminous objects *e.g.* a lighted candle/torch/bulb, the Sun, stars. The moon is not a luminous object! Why?

Most objects however are not luminous, objects which do not give out their own light. We can see them because they reflect light into our eyes.

3.4 Reflection

3.4.1 Diffuse reflection

(NOTE TO SELF: diag of light rays hitting a rough surface) (NOTE TO SELF: diag of light rays hitting a polished surface)

Most objects do not have perfectly smooth surfaces. Because of this different parts of the surface reflect light rays in different directions (angles).

3.4.2 Regular reflection

Mirrors and highly polished surfaces give regular reflection. A mirror is a piece of glass with a thin layer of silver (aluminium is commonly used) on the rear surface

Light is reflected according to the laws of reflection.

3.4.3 Laws of reflection

(NOTE TO SELF: diag of i N r rays striking a mirror define i r N)

Laws of reflection:
The angle of incidence is **always** equal to the angle of reflection
The incident ray, the normal and the reflected ray are all in the same plane

[1] There are some substances where light moves so slowly you can walk faster than it moves - more on this later!

Note that the angle of incidence i is between the incident ray and the normal[2] to the surface, **not** between the incident ray and the surface of the mirror.

There are two forms of an image formed by reflection: *real and virtual.*

A **real** image is formed by the **actual** intersection of light rays. It is always inverted (upside down) and can be formed on a screen

A **virtual** image is formed by the **apparent** intersection of light rays. It is always erect and cannot be formed on a screen.

Images formed in plane mirrors are virtual images. Real images are formed by lenses (*e.g.* the image on a cinema screen) or by curved mirrors.

3.4.4 Lateral inversion

Where some thing is back to front *e.g.* AMBULANCE

(NOTE TO SELF: How a periscope works, why images are right!)

3.5 Curved Mirrors

3.5.1 Concave Mirrors (Converging Mirrors)

(NOTE TO SELF: how to remember a con_cave from a convex mirror)

(NOTE TO SELF: diag of light rays striking a concave mirror base ray diag! define things)

P = pole of mirror F = focal point (focus) C = centre of curvature —CP— = radius of curvature (r) —FP— = focal length (f)

r = 2f

Rules for ray tracing:

Rays of light that arrive parallel to principal axis leave the surface of the mirror through the focal point
Rays of light that arrive through the focal point leave the surface of the mirror parallel to principal axis
Rays of light that arrive through the centre of curvature leave the surface back through the centre of curvature

(NOTE TO SELF: Maybe diag for each of these above!)

In order to obtain the position, nature (real or virtual) and size of the image we need just apply the rules above. The details depend on the distance of the object from the mirror surface..

Object outside C

diag

Image - located between c and f - inverted (upside down) - diminished (reduced in size, smaller) - real

Object at C

diag

Image - located at c - inverted (upside down) - same size - real

Object between C and f

diag

Image - located outside c - inverted (upside down) - magnified (increased in size, larger) - real

Object at f

diag

[2] The normal is a line that is perpendicular to the surface *i.e.* the angle between the line and the surface is 90°

Image - located at infinity

Object between f and P

diag

Image - located behind the mirror - erect (right side up) - magnified (increased in size, larger) - virtual

3.5.2 Convex Mirrors

diag base ray diag! define things

Image - located behind the mirror - erect (right side up) - diminished (reduced in size, smaller) - virtual

We can also arrive at the position and nature of the image by calculation, using the following formula

1/u + 1/v = 1/f

AND

Magnification (=M)

m = v/u

(We will deal with the 'proff' for these later)

N.B.

distance between f and P = f (focal length negative for convex mirror) distance between o and P = u (object distance) distance between i and P = v (image distance negative for virtual image)

concave mirror f is positive convex mirror f is negative real image v is positive virtual image v is negative

Lots of examples with numbers

e.g. u = 10cm v = 20cm m=2

things are bigger if m ¿ 1 (and -1) things are smaller if 0 ¿ m ¿ 1 (and -1)

Uses of Convex mirrors

- image is always erect
- wide range of view

They often used in shops, double decker busses, dangerous bends in roads, wing mirrors of cars

Disadvantage False sense of distance (objects seem closer than they actually are)

uses of concave mirrors

They are usually used as make-up mirrors or shaving mirrors and in reflecting telescopes

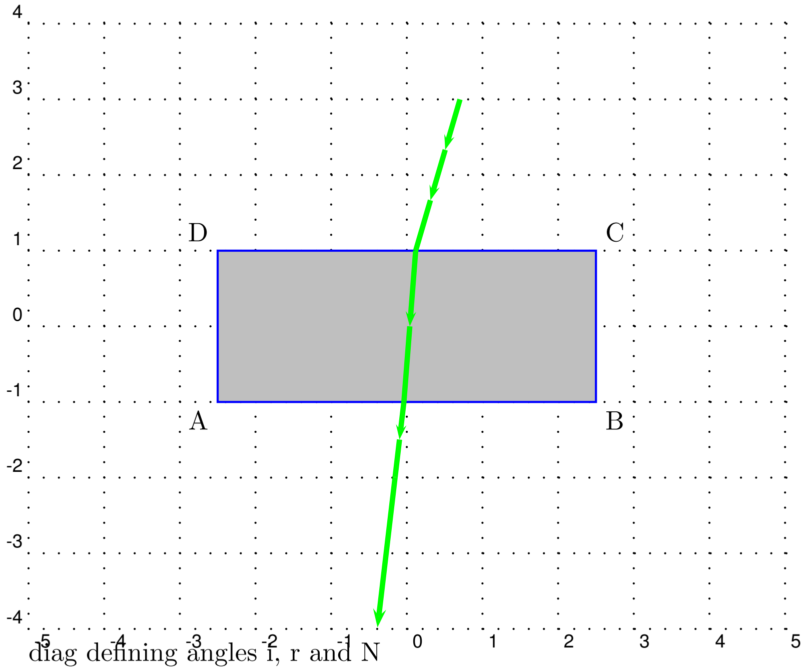
Advantages

- if object is inside f the image is magnified

MAYBE MENTION 'Virtual object' in convex mirror for completeness

3.5.3 Refraction

When light travels from one medium to another it changes direction, except when it is incident normally on the separating surface. The change of direction is caused by the change in the velocity of the light as it passes from one medium to the other.



diag defining angles i, r and N

Light is refracted according to the laws of refraction:

3.5.4 Laws of Refraction

Laws of Refraction:
sin i /sin r is a constant for two given media (Snell's Law)
The incident ray, the normal and the refracted ray are all in the same plane

Sin i /sin r is known as the refractive index (n)

EXP verify Snell's law

EXP coin in an empty cup, move head till it disappears. Then fill with water. Can you see it?

diag real apparent depth

Due to refraction a body which is at O beneath the water appears to be at I when seen from above. As a result of refraction the pool appears to be 1.5m instead of 2m. The relationship between the real depth and the apparent depth is determined by the refractive index of water, which is 4/3

diag of water depth in pool

n = sin i /sin r =—AP—/—PI— / —AP—/—PO— —PO—/—PI—

However the above diagram is greatly exaggerated in size. In practise

—PO—   —AO— and —PI—   —AI—

So n =—AO—/—AI— == real/apparent depth!

More than two media

$1\_n\_3 = 1\_n\_2 \times 2\_n\_3$

e.g. air glass water

The refractive index of glass is 3/2 and the refractive index of water is 4/3. The refractive index from water to glass

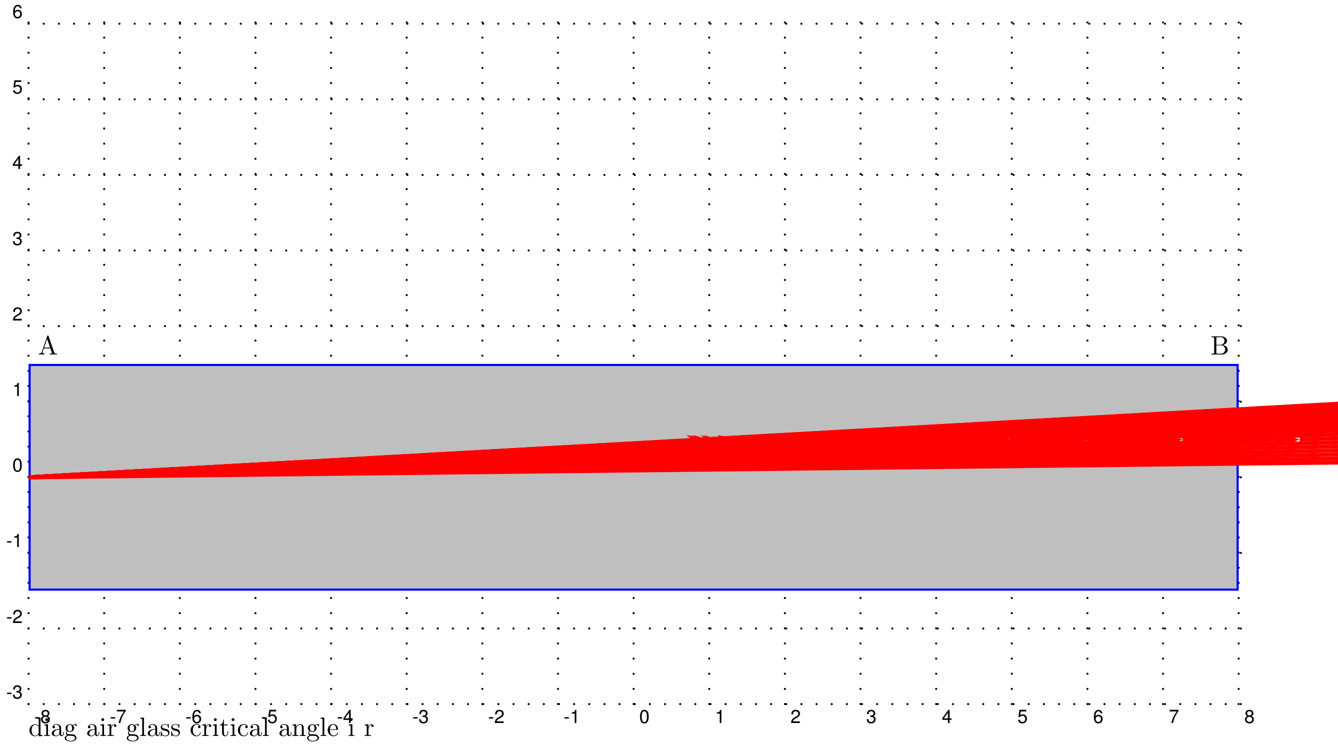
a_n_w= 4/3 =¿ w_n_a =3/4

$w\_n\_g = w\_n\_a \times a\_n\_g = 3/4 \times 3/2 = 9/8$

3.5.5 Total Internal Reflection

This can only happen when light is travelling into a less dense medium

Refrectionnumbers are $n_1 = 2$ and $n_2 = 1$:



diag air glass critical angle i r

When light travels from glass to air it is refracted away from the normal as in a above. As the angle of incidence is increased the angle of refraction eventually reaches 90° as in b If the angle of incidence is increased further beyond this value total internal reflection occurs as in c

The critical angle c is the angle of incidence corresponding to the angle of refraction of 90°

From the above the refractive index from from air to glass

n=sin90/sin c = 1/sin c

Also sin c = 1 / n = 2/3 in the case of glass.

Sin c = 0.6667 =¿ critical angle for glass is $41°49'$

Total internal reflection has some very useful properties eg

diag prism for turning light 180°

daig prism for turning light 90°

The size of the final image in a pair of binoculars depends on the distance travelled by the light **within** the binoculars. By using two prisms this distance can be increased without increasing the length of the binoculars.

Fibre Optics

Water can be directed from one place to another by confining it within a pipe. In the same way light can be directed from one place to another by confining it within a single glass fibre. The light is kept within the fibre by total internal reflection. The amount of light which can be carried by a single fibre is very small so it is usual to form a light tube tapping a few thousand fibres together. On great advantage of such a light tube is flexibility; it can be ties in knots and still function. However since total internal reflection only occurs when light is going from a medium to a less dense medium, it is necessary ti coat each fibre with glass of a lower refractive index. Otherwise light would leak from one fibre at their points of contact.

Light tube can be used to bring light from a lamp to an object, thus illuminating the object. A second light tube can then used to carry light from the illuminated object to an observer, thus enabling the object to be seen and photographed. The procedure has been used to photograph the digestive system the reproductive system and many other parts of the human body. In the case of the light tube carrying light from the object to the observer, it is vital that the individual fibres in the tube do not cross each other, otherwise the image will become garbled. Like radio waves, light waves are electromagnetic. (cf section em) However, their shorter wavelength and higher frequency means that a single light beam can carry far more telephone conservations at one time compared with a radio wave.

In the case of long fibre cables it would be necessary to incorporate a device to boost the intensity of the light to make up for losses due to absorption. Nevertheless the system has great potential for the communication industry, including the possibility of transmitting pictures over long distances.

As we said earlier, the reason why light bends when going from one medium to another is because of the change of velocity. This will be dealt with in more detail later. For the moment we consider only the implication for the refractive index.

1_n_2 = velocity medium 1 / velocity medium 2

3.5.6 Mirage

Yet another effect of refraction is the mirage. The most common mirage occurs in warm weather when motorists see what appears to be a pool of water on the road close to the horizon. The explanation is this: when air is heated it expands and becomes less dense and when it cools it contracts and becomes more dense. In summer the ground is hot and the layer of air nearest the ground is hot. The layer above that is cooler, etc etc. A of light coming from the blue sky passes down through the layers of air which are getting progressively less dense. As it does so it is progressively bent, as shown in the diagram. As a result an image of the blue sky is seen on the road and is taken to be a pool of water!

3.6 The Electromagnetic Spectrum

They are transverse waves. They can travel through empty space. They travel at the speed of light.

Short wavelength Long wavelength

Dangers

Microwaves kill living cells Ultraviolet light causes skin cancer and can kill living cells

Uses

Radio waves are used in TV's and radios Microwaves are used to heat foods, they are also used to transmit signals to satellites and back to earth. Infrared light is used to asses heat loss from buildings, weather forecasting, locating survivors in earthquakes and in TV remote controls. Visible light is used in fibre optics. Ultraviolet light produces vitamins in the skin if it is in small doses. Suntan. X-Rays are used to look at bones inside your body, my body and everybody's body. g-rays are used to treat cancer.

3.7 Important Equations and Quantities

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
			or	

Table 3.1: Units used in **Optics**

Chapter 4

Vectors

(NOTE TO SELF: SW: initially this chapter had a very mathematical approach. I have toned this down and tried to present in a logical way the techniques of vector manipulation after first exploring the mathematical properties of vectors. Most of the PGCE comments revolved around the ommision of the graphical techniques of vector addition (i.e. scale diagrams), incline questions and equilibrium of forces. I have addressed the first two; equilibrium of forces and the triangle law of three forces in equilibrium I think would be better off in the Forces Chapter. Also the Forces chapter should include some examples of incline planes. Inclines are introduced here merely as an example of components in action!)

4.1 PGCE Comments

- Use a+b = c to illustrate that a = c-b under subtraction.
- Summarise properties of vectors at end.
- Mention that vector multiplication exists in two forms- neither introduced at school, but used implicitly in work and motor rules.

4.2 'TO DO' LIST

- Mention that the direction of a vector is measured from its tail.
- Also include explanations of the ways in which to specify direction (e.g. bearing, compass points, angle from vertical, etc.)
- Include exam-type questions involving flowing rivers or aeroplanes in wind. These questions link displacement, velocity and time and also test the learners understanding of vectors combining to give a combined result (i.e. resultant).
- Mention that displacement at a point is the directed line from the start to that point.
- Rewrite section on velocity to make clear the distinction between average and instantaneous rates of change (velocity and speed). The Δ's in the equations imply we are calculating average quantities. Mention that we take the limit of a small time interval to give instantaneous quantities. Perhaps the example of a parabola with average gradient and gradient of

tangent can be used as an illustration. Else defer until chapter on Graphs and Equations of Motion. Instantaneous velocity: reading on the speedometer in a direction tangent to the path. Instantaneous speed is magnitude of instantaneous velocity but average speed is not equal to magnitude of average velocity. Average speed and average velocity are the total distance and resultant displacement over the time interval related to that part of the path. The example of the circular track uses these definitions and is an important illustration of the differences. Instantaneous calculated at a certain instant in time while average is calculated over an interval.

- Include relative velocity.
- Address PGCE comments above and comments in the text.

4.3 Introduction

"A vector is 'something' that has both magnitude and direction." "'Things'? What sorts of 'things'?" Any piece of information which contains a magnitude and a related direction can be a vector. A vector should tell you *how much* and *which way*.

Consider a man driving his car east along a highway at 100 km/h. What we have given here is a vector– the car's velocity. The car is moving at 100 km/h (this is the magnitude) and we know where it is going– east (this is the direction). Thus, we know the speed and direction of the car. These two quantities, a magnitude and a direction, form a vector we call velocity.

Definition: A *vector* is a measurement which has both magnitude and direction.

In physics magnitudes often have directions associated with them. If you push something it is not very useful knowing just how hard you pushed. A direction is needed too. Directions are extremely important, especially when dealing with situations more complicated than simple pushes and pulls.

Different people like to write vectors in different ways. Anyway of writing a vector so that it has both magnitude and direction is valid.

Are vectors physics? No, vectors themselves are not physics. Physics is just a description of the world around us. To describe something we need to use a language. The most common language used to describe physics is mathematics. Vectors form a very important part of the mathematical description of physics, so much so that it is absolutely essential to master the use of vectors.

4.3.1 Mathematical representation

Numerous notations are commonly used to denote vectors. In this text, vectors will be denoted by symbols capped with an arrow. As an example, $\overrightarrow{s}$, $\overrightarrow{v}$ and $\overrightarrow{F}$ are all vectors (they have both magnitude and direction). Sometimes just the magnitude of a vector is required. In this case, the arrow is ommitted. In other words, F denotes the magnitude of vector $\overrightarrow{F}$. $|\overrightarrow{F}|$ is another way of representing the size of a vector.

4.3.2 Graphical representation

Graphically vectors are drawn as arrows. An arrow has both a magnitude (how long it is) and a direction (the direction in which it points). For this reason, arrows are vectors.

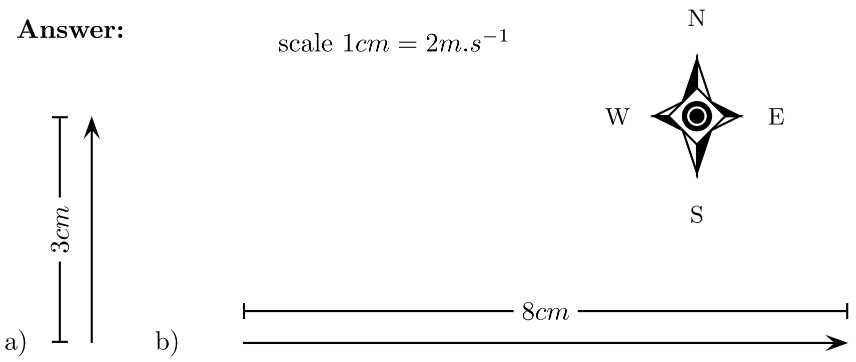
In order to draw a vector accurately we must specify a scale and include a reference direction in the diagram. A scale allows us to translate the length of the arrow into the vector's magnitude. For instance if one chose a scale of $1cm = 2N$ ($1cm$ represents $2N$), a force of magnitude $20N$ would be represented as an arrow $10cm$ long. A reference direction may be a line representing a horizontal surface or the points of a compass.

Worked Example 3

Drawing vectors

Question: Using a scale of $1cm = 2m.s^{-1}$ represent the following velocities:
a) $6m.s^{-1}$ north
b) $16m.s^{-1}$ east

Answer:



4.4 Some Examples of Vectors

4.4.1 Displacement

Imagine you walked from your house to the shops along a winding path through the veld. Your route is shown in blue in Figure 12.3. Your sister also walked from the house to the shops, but she decided to walk along the pavements. Her path is shown in red and consisted of two straight stretches, one after the other. Although you took very different routes, both you and your sister walked from the house to the shops. The overall effect was the same! Clearly the shortest path from your house to the shops is along the straight line between these two points. The length of this line and the direction from the start point (the house) to the end point (the shops) forms a very special vector known as **displacement**. Displacement is assigned the symbol $\overrightarrow{s}$.

Definition: *Displacement* is defined as the magnitude and direction of the straight line joining one's starting point to one's final point.

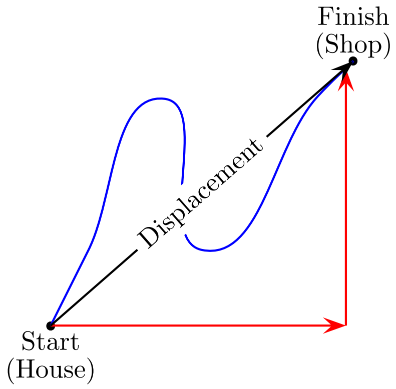


Figure 4.1: Illustration of Displacement

OR

Definition: *Displacement* is a vector with direction pointing from some initial (starting) point to some final (end) point and whose magnitude is the straight-line distance from the starting point to the end point.

(NOTE TO SELF: choose one of the above)

In this example both you and your sister had the same displacement. This is shown as the black arrow in Figure 12.3. Remember displacement is not concerned with the actual path taken. It is only concerned with your start and end points. It tells you the length of the straight-line path between your start and end points and the direction from start to finish. The distance travelled is the length of the path followed and is a scalar (just a number). Note that the magnitude of the displacement need not be the same as the distance travelled. In this case the magnitude of your displacement would be considerably less than the actual length of the path you followed through the veld!

4.4.2 Velocity

Definition: *Velocity* is the rate of change of displacement with respect to time.

The terms *rate of change* and *with respect to* are ones we will use often and it is important that you understand what they mean. Velocity describes how much displacement changes for a certain change in time.

We usually denote a change in something with the symbol Δ (the Greek letter Delta). You have probably seen this before in maths– the gradient of a straight line is $\frac{\Delta y}{\Delta x}$. The gradient is just how much y changes for a certain change in x. In other words it is just the rate of change of y with respect to x. This means that velocity must be

$$\overrightarrow{v} = \frac{\Delta \overrightarrow{s}}{\Delta t} = \frac{\overrightarrow{s}_{final} - \overrightarrow{s}_{initial}}{t_{final} - t_{initial}}. \tag{4.1}$$

(NOTE TO SELF: This is actually average velocity. For instantaneous Δ's change to differentials. Explain that if Δ is large then we have average velocity else for infinitesimal time interval instantaneous!)

What then is speed? Speed is how quickly something is moving. How is it different from velocity? Speed is not a vector. It does not tell you which direction something is moving, only how fast. Speed is the magnitude of the velocity vector (NOTE TO SELF: instantaneous speed is the magnitude of the instantaneous velocity.... not true of averages!).

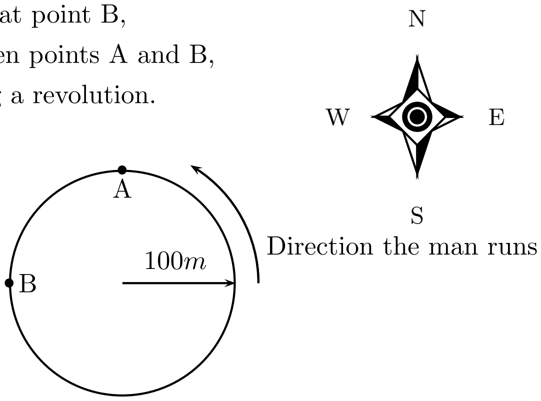
Consider the following example to test your understanding of the differences between velocity and speed.

Worked Example 4

Speed and Velocity

Question: A man runs around a circular track of radius $100m$. It takes him $120s$ to complete a revolution of the track. If he runs at constant speed calculate:

1. his speed,
2. his instantaneous velocity at point A,
3. his instantaneous velocity at point B,
4. his average velocity between points A and B,
5. his average velocity during a revolution.



Answer:

1. To determine the man's speed we need to know the distance he travels and how long it takes. We know it takes $120s$ to complete one revolution of the track.

   Step 1 : First we find the distance the man travels

   What distance is one revolution of the track? We know the track is a circle and we know its radius, so we can determine the perimeter or distance around the circle. We start with the equation for the circumference of a circle

$$\begin{aligned} C &= 2\pi r \\ &= 2\pi(100m) \\ &= 628.3\ m. \end{aligned}$$

   So we know the distance the man covers in one revolution is $628.3\ m$.

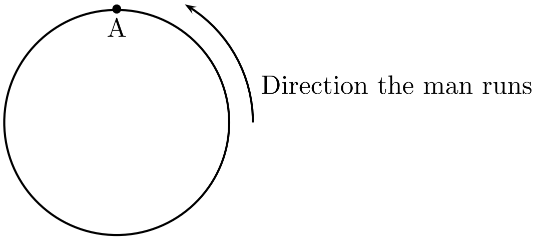
Step 2 : Now we determine speed from the distance and time.

We know that speed is distance covered per unit time. So if we divide the distance covered by the time it took we will know how much distance was covered for every unit of time.

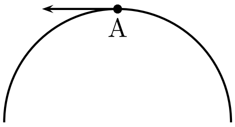
$$\begin{aligned} v &= \frac{Distance\ travelled}{time\ taken} \\ &= \frac{628.3m}{120s} \\ &= 5.23\ m.s^{-1} \end{aligned}$$

2. *Step 3 : Determine his instantaneous velocity at A*

   Consider the point A in the diagram.



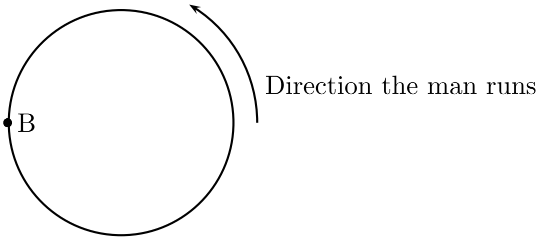
   We know which way the man is running around the track and we know his speed. His velocity at point A will be his speed (the magnitude of the velocity) plus his direction of motion (the direction of his velocity). The instant that he arrives at A he is moving as indicated in the diagram below.



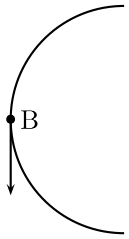
   So his velocity vector will be 5.23 $m.s^{-1}$ West.

3. *Step 4 : Determine his instantaneous velocity at B*

   Consider the point B in the diagram.



We know which way the man is running around the track and we know his speed. His velocity at point B will be his speed (the magnitude of the velocity) plus his direction of motion (the direction of his velocity). The instant that he arrives at B he is moving as indicated in the diagram below.



So his velocity vector will be 5.23 $m.s^{-1}$ South.

4. *Step 5 : Now we determine his average velocity between A and B*
(NOTE TO SELF: add this here to further stress the difference between average and instantaneous velocities, as well as the difference between magnitude of average velocity and average speed!)
5. *Step 6 : Now we calculate his average velocity over a complete revolution.*
The definition of average velocity is given earlier and requires that you know the total displacement and the total time. The total displacement for a revolution is given by the vector from the initial point to the final point. If the man runs in a circle then he ends where he started. This means the vector from his initial point to his final point has zero length.
So a calculation of his average velocity follows:

$$\begin{aligned} \overrightarrow{v} &= \frac{\Delta \overrightarrow{s}}{\Delta t} \\ &= \frac{0m}{120s} \\ &= 0 \; m.s^{-1} \end{aligned}$$

Remember - displacement can be zero even when distance is not!

4.4.3 Acceleration

Definition: *Acceleration* is the rate of change of velocity with respect to time.

Acceleration is also a vector. Remember that velocity was *the rate of change of displacement with respect to time* so we expect the velocity and acceleration equations to look very similar. In fact,

$$\overrightarrow{a} = \frac{\Delta \overrightarrow{v}}{\Delta t} = \frac{\overrightarrow{v}_{final} - \overrightarrow{v}_{initial}}{t_{final} - t_{initial}} \tag{4.2}$$

(NOTE TO SELF: average and instantaneous distiction again! expand further- what does it mean.)

Acceleration will become very important later when we consider forces.

4.5 Mathematical Properties of Vectors

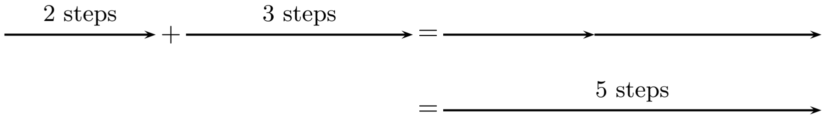
Vectors are mathematical objects and we will use them to describe physics in the language of mathematics. However, first we need to understand the mathematical properties of vectors (e.g. how they add and subtract).

We will now use arrows representing displacements to illustrate the properties of vectors. Remember that displacement is just one example of a vector. We could just as well have decided to use forces to illustrate the properties of vectors.

Using vectors i
an important
skill you **MUS**
master!

4.5.1 Addition of Vectors

If we define a displacement vector as 2 steps in the forward direction and another as 3 steps in the forward direction then adding them together would mean moving a total of 5 steps in the forward direction. Graphically this can be seen by first following the first vector two steps forward and then following the second one three steps forward:



We add the second vector at the end of the first vector, since this is where we now are after the first vector has acted. The vector from the tail of the first vector (the starting point) to the head of the last (the end point) is then the sum of the vectors. This is the tail-to-head method of vector addition.

As you can convince yourself, the order in which you add vectors does not matter. In the example above, if you decided to first go 3 steps forward and then another 2 steps forward, the end result would still be 5 steps forward.

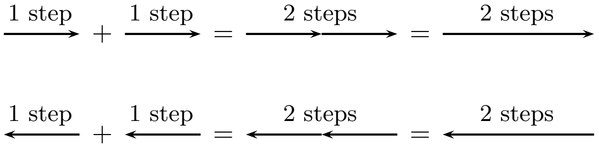
The final answer when adding vectors is called the *resultant*.

Definition: The *resultant* of a number of vectors is the single vector whose effect is the same as the individual vectors acting together.

In other words, the individual vectors can be replaced by the resultant– the overall effect is the same. If vectors $\overrightarrow{a}$ and $\overrightarrow{b}$ have a resultant $\overrightarrow{R}$, this can be represented mathematically as,

$$\overrightarrow{R} \quad = \quad \overrightarrow{a} + \overrightarrow{b}.$$

Let us consider some more examples of vector addition using displacements. The arrows tell you how far to move and in what direction. Arrows to the right correspond to steps forward, while arrows to the left correspond to steps backward. Look at all of the examples below and check them.



Let us test the first one. It says one step forward and then another step forward is the same as an arrow twice as long– two steps forward.

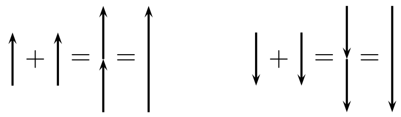
It is possible that you end up back where you started. In this case the net result of what you have done is that you have gone nowhere (your start and end points are at the same place). In this case, your resultant displacement is a vector with length zero units. We use the symbol $\overrightarrow{0}$ to denote such a vector:

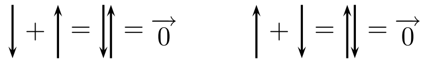
$$\xrightarrow{\text{1 step}} + \xleftarrow{\text{1 step}} = \underset{\text{1 step}}{\overset{\text{1 step}}{\rightleftarrows}} = \overrightarrow{0}$$

$$\xleftarrow{\text{1 step}} + \xrightarrow{\text{1 step}} = \underset{\text{1 step}}{\overset{\text{1 step}}{\leftrightarrows}} = \overrightarrow{0}$$

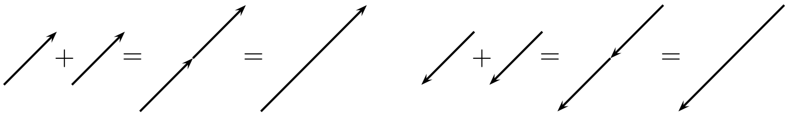
Check the following examples in the same way. Arrows up the page can be seen as steps left and arrows down the page as steps right.

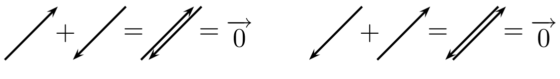
Try a couple to convince yourself!



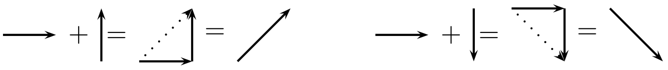


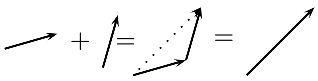
It is important to realise that the directions aren't special– 'forward and backwards' or 'left and right' are treated in the same way. The same is true of any set of parallel directions:





In the above examples the separate displacements were parallel to one another. However the same 'tail-to-head' technique of vector addition can be applied to vectors in any direction.



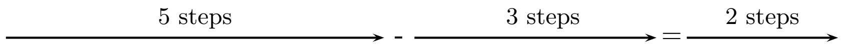


Now you have discovered one use for vectors; describing resultant displacement– how far and in what direction you have travelled after a series of movements.

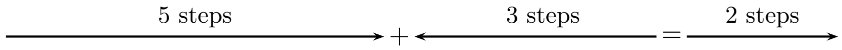
Although vector addition here has been demonstrated with displacements, all vectors behave in exactly the same way. Thus, if given a number of forces acting on a body you can use the same method to determine the resultant force acting on the body. We will return to vector addition in more detail later.

4.5.2 Subtraction of Vectors

What does it mean to subtract a vector? Well this is really simple; if we have 5 apples and we subtract 3 apples, we have only 2 apples left. Now lets work in steps; if we take 5 steps forward and then subtract 3 steps forward we are left with only two steps forward:

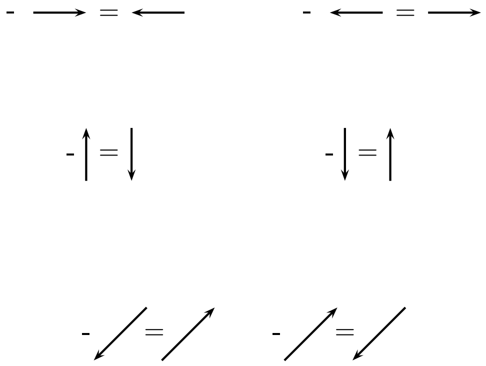


What have we done? You originally took 5 steps forward but then you took 3 steps back. That backward displacement would be represented by an arrow pointing to the left (backwards) with length 3. The net result of adding these two vectors is 2 steps forward:



Thus, subtracting a vector from another is the same as adding a vector in the opposite direction (i.e. subtracting 3 steps forwards is the same as adding 3 steps backwards).

This suggests that in this problem arrows to the right are positive and arrows to the left are negative. More generally, vectors in opposite directions differ in sign (i.e. if we define up as positive, then vectors acting down are negative). Thus, changing the sign of a vector simply reverses its direction:



In mathematical form, subtracting $\overrightarrow{a}$ from $\overrightarrow{b}$ gives a new vector $\overrightarrow{c}$:

$$\begin{aligned}\overrightarrow{c} &= \overrightarrow{b} - \overrightarrow{a} \\ &= \overrightarrow{b} + (-\overrightarrow{a})\end{aligned}$$

This clearly shows that subtracting vector $\overrightarrow{a}$ from $\overrightarrow{b}$ is the same as adding $(-\overrightarrow{a})$ to $\overrightarrow{b}$. Look at the following examples of vector subtraction.

$$\longrightarrow \text{ - } \longrightarrow = \longrightarrow + \longleftarrow = \overrightarrow{0}$$

$$\longrightarrow \text{ - } \longleftarrow = \longrightarrow + \longrightarrow = \xrightarrow{\qquad\quad}$$

4.5.3 Scalar Multiplication

What happens when you multiply a vector by a scalar (an ordinary number)?

Going back to normal multiplication we know that 2×2 is just 2 groups of 2 added together to give 4. We can adopt a similar approach to understand how vector multiplication works.

$$2 \text{ x } \longrightarrow = \longrightarrow + \longrightarrow = \xrightarrow{\qquad\quad}$$

4.6 Techniques of Vector Addition

Now that you have been acquainted with the mathematical properties of vectors, we return to vector addition in more detail. There are a number of techniques of vector addition. These techniques fall into two main categories- graphical and algebraic techniques.

4.6.1 Graphical Techniques

Graphical techniques involve drawing accurate scale diagrams to denote individual vectors and their resultants. We next discuss the two primary graphical techniques, the tail-to-head technique and the paralelogram method.

The Tail-to-head Method

In describing the mathematical properties of vectors we used displacements and the tail-to-head graphical method of vector addition as an illustration. In the tail-to-head method of vector addition the following strategy is followed:

- Choose a scale and include a reference direction.
- Choose any of the vectors to be summed and draw it as an arrow in the correct direction and of the correct length– remember to put an arrowhead on the end to denote its direction.

- Take the next vector and draw it as an arrow starting from the arrowhead of the first vector in the correct direction and of the correct length.
- Continue until you have drawn each vector– each time starting from the head of the previous vector. In this way, the vectors to be added are drawn one after the other tail-to-head.
- The resultant is then the vector drawn from the tail of the first vector to the head of the last. Its magnitude can be determined from the length of its arrow using the scale. Its direction too can be determined from the scale diagram.

Worked Example 5

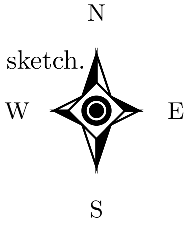
Tail-to-Head Graphical Addition I

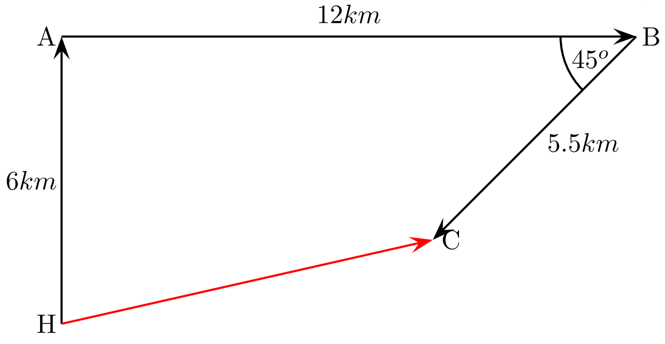
Question: A ship leaves harbour H and sails $6km$ north to port A. From here the ship travels $12km$ east to port B, before sailing $5.5km$ south-west to port C. Determine the ship's resultant displacement using the tail-to-head technique of vector addition.

Answer:

Now, we are faced with a practical issue: in this problem the displacements are too large to draw them their actual length! Drawing a $2km$ long arrow would require a very big book. Just like cartographers (people who draw maps), we have to choose a scale. The choice of scale depends on the actual question– you should choose a scale such that your vector diagram fits the page. Before choosing a scale one should always draw a rough sketch of the problem. In a rough sketch one is interested in the approximate shape of the vector diagram.

Step 1 : Draw a rough sketch of the situation
Its easy to understand the problem if we first draw a quick sketch.





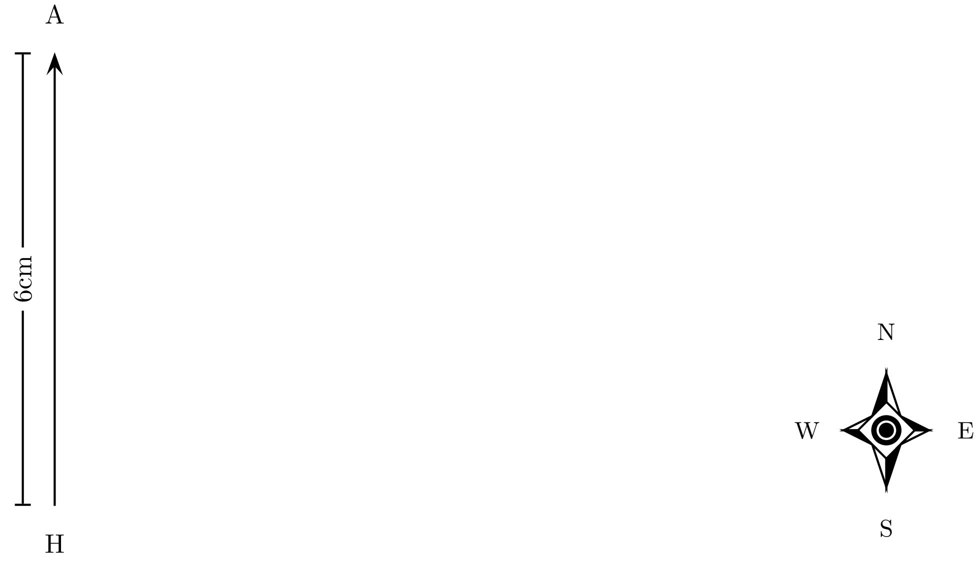
In a rough sketch one should include all of the information given in the problem. All of the magnitudes of the displacements are shown and a compass has been included as a reference direction.

Step 2 : Next we choose a scale for our vector diagram

It is clear from the rough sketch that choosing a scale where $1cm$ represents $1km$ (scale: $1cm = 1km$) would be a good choice in this problem)– the diagram will then take up a good fraction of an A4 page. We now start the accurate construction.

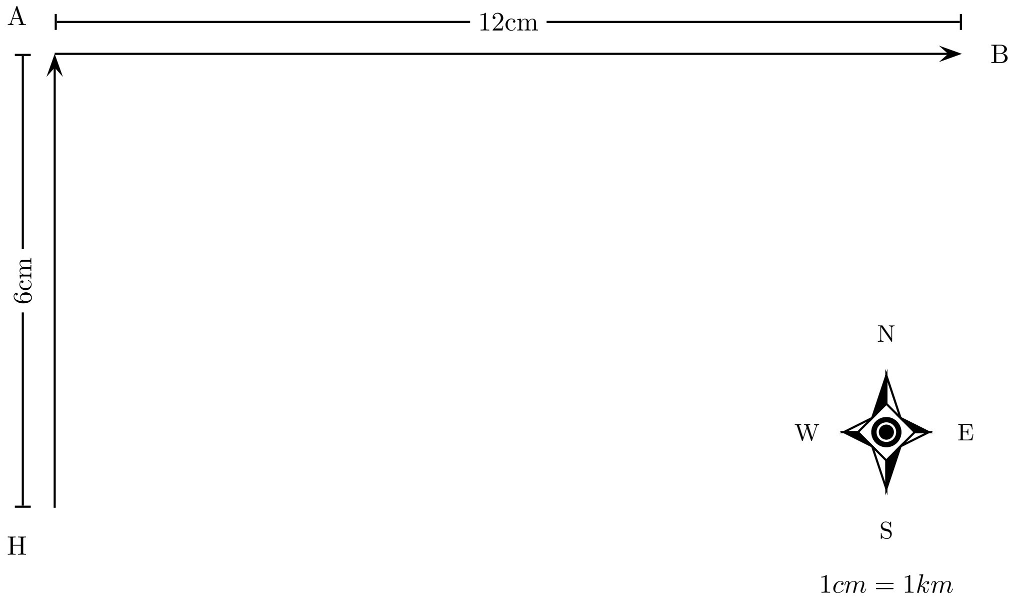
Step 3 : Now we construct our scaled vector diagram

Contruction Step 1: Starting at the harbour H we draw the first vector $6cm$ long in the direction north (remember in the diagram $1cm$ represents $1km$):

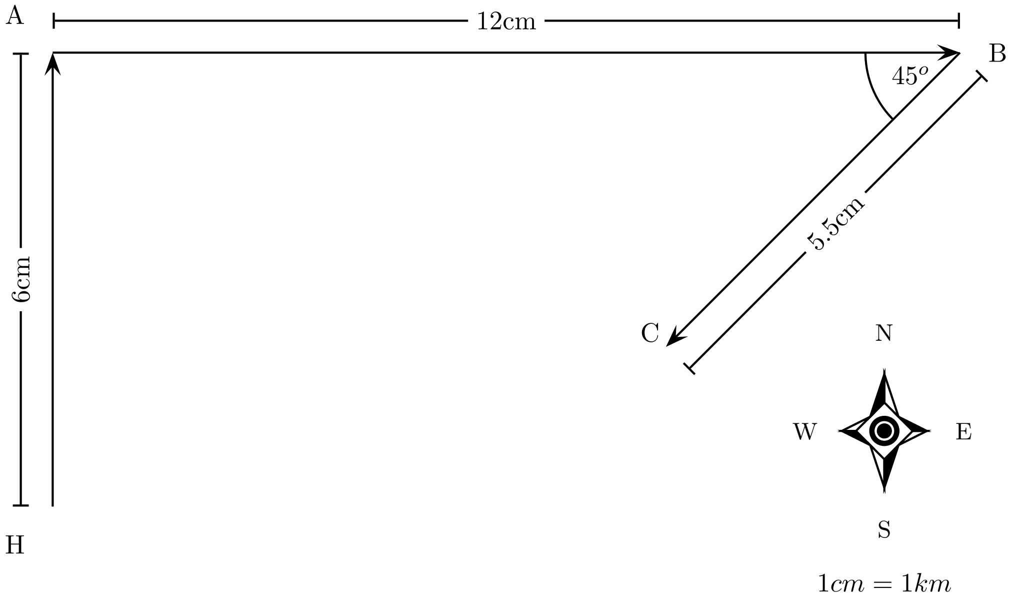


$1cm = 1km$

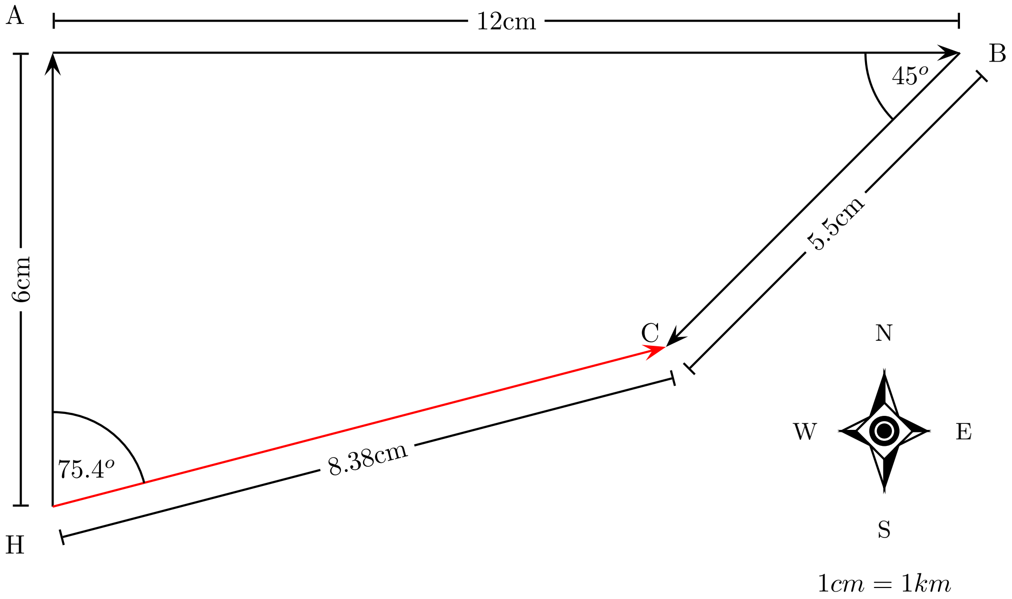
Construction Step 2: Since the ship is now at port A we draw the second vector $12cm$ long starting from this point in the direction east:


A
12cm
B
6cm
H
N
W
E
S
1cm = 1km

Construction Step 3: Since the ship is now at port B we draw the third vector $5.5cm$ long starting from this point in the direction south-west. A protractor is required to measure the angle of 45^o.



Construction Step 4: As a final step we draw the resultant displacement from the starting point (the harbour H) to the end point (port C). We use a ruler to measure the length of this arrow and a protractor to determine its direction

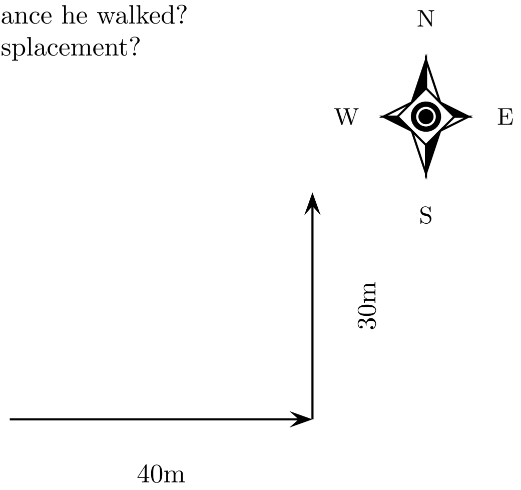


Step 4 : Apply the scale conversion
We now use the scale to convert the length of the resultant in the scale diagram to the actual displacement in the problem. Since we have chosen a scale of $1cm = 1km$ in this problem the resultant has a magnitude of 8.38 km. The direction can be specified in terms of the angle measured either as 75.4^o east of north or on a bearing of 75.4^o.
Step 5 : Quote the final answer
The resultant displacement of the ship is 8.38 km on a bearing of 75.4^o!

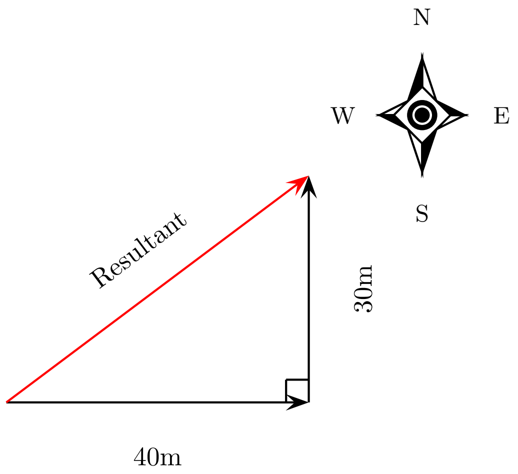
Worked Example 6

Tail-to-Head Graphical Addition II

Question: A man walks 40 m East, then 30 m North.
a) What was the total distance he walked?
b) What is his resultant displacement?



Answer:
Step 1 : a) Determine the distance that the man traveled
In the first part of his journey he traveled 40 m and in the second part he traveled 30 m. This gives us a total distance traveled of $40 + 30 = 70$ m.
Step 2 : b) Determine his resultant displacement - start by drawing a rough sketch
The man's resultant displacement is the **vector** from where he started to where he ended. It is the sum of his two separate displacements. We will use the tail-to-head method of accurate construction to find this vector. Here is our rough sketch:

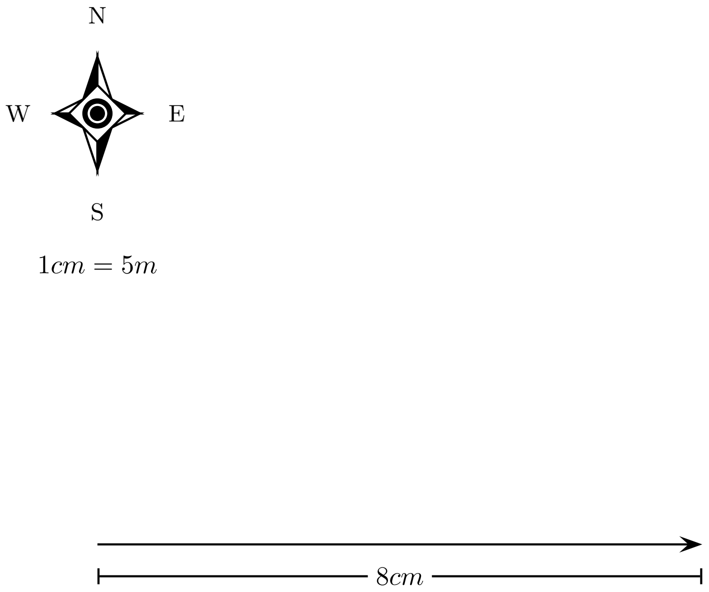


Step 3 : Choose a suitable scale

A scale of $1cm$ represents $5m$ ($1cm = 5m$) is a good choice here. Now we can begin the process of construction.

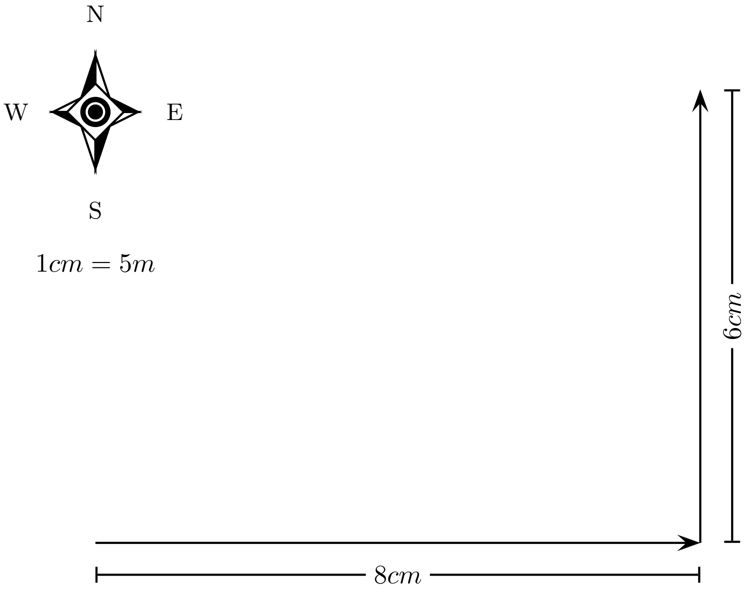
Step 4 : Draw the first vector according to scale

We draw the first displacement as an arrow $8cm$ long (according to the scale $8cm = 8 \times 5m = 40m$) in the direction east:

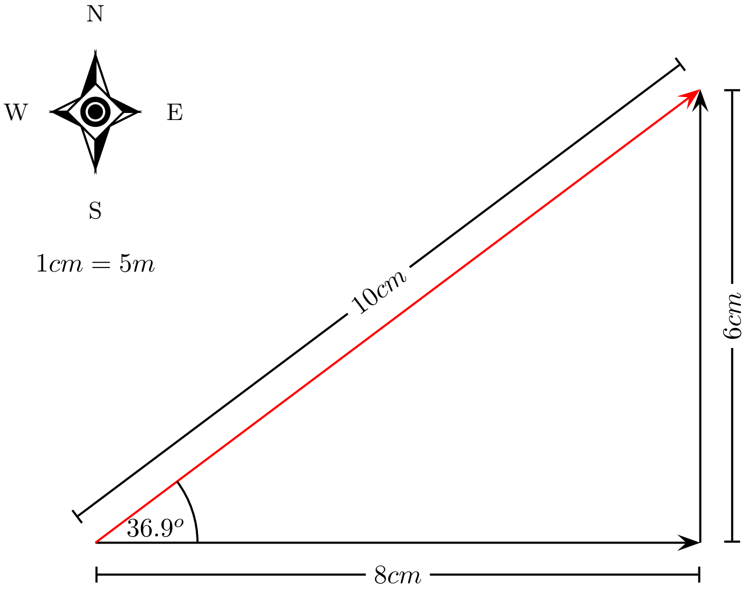


Step 5 : Draw the second vector according to scale

Starting from the head of the first vector we draw the second displacement as an arrow $6cm$ long (according to the scale $6cm = 6 \times 5m = 30m$) in the direction north:



Step 6 : Determine the resultant vector
Now we connect the starting point to the end point and measure the length and direction of this arrow (the resultant)



Step 7 : Apply the scale conversion
Finally we use the scale to convert the length of the resultant in the scale diagram to the actual magnitude of the resultant displacement. According to the chosen scale $1cm = 5m$. Therefore $10cm$ represents $50m$. The resultant displacement is then $50m$ 36.9^o north of east.

The Parallelogram Method

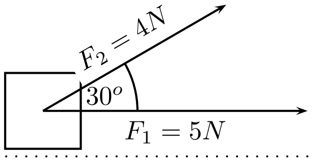
When needing to find the resultant of two vectors another graphical technique can be applied- the parallelogram method. The following strategy is employed:

- Choose a scale and a reference direction.
- Choose either of the vectors to be added and draw it as an arrow of the correct length in the correct direction.
- Draw the second vector as an arrow of the correct length in the correct direction **from the tail of the first vector**.
- Complete the parallelogram formed by these two vectors.
- The resultant is then the diagonal of the parallelogram. Its magnitude can be determined from the length of its arrow using the scale. Its direction too can be determined from the scale diagram.

Worked Example 7

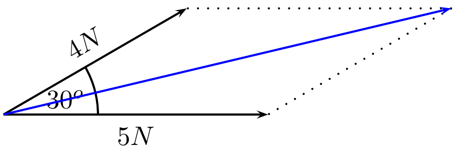
Parallelogram Method of Graphical Addition I

Question: A force of $F_1 = 5N$ is applied to a block in a horizontal direction. A second force $F_2 = 4N$ is applied to the object at an angle of 30^o above the horizontal.



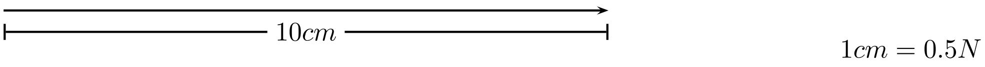
Determine the resultant force acting on the block using the parallelogram method of accurate construction.

Answer:
Step 1 : Firstly make a rough sketch of the vector diagram

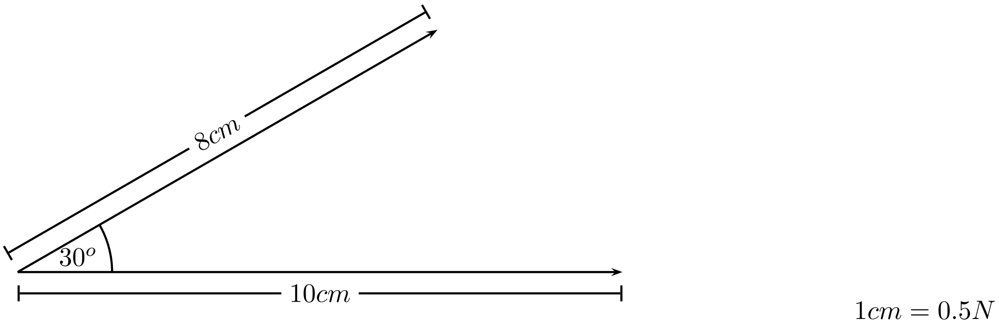


Step 2 : Choose a suitable scale
In this problem a scale of $1cm = 0.5N$ would be appropriate, since then the vector diagram would take up a reasonable fraction of the page. We can now begin the accurate scale diagram.

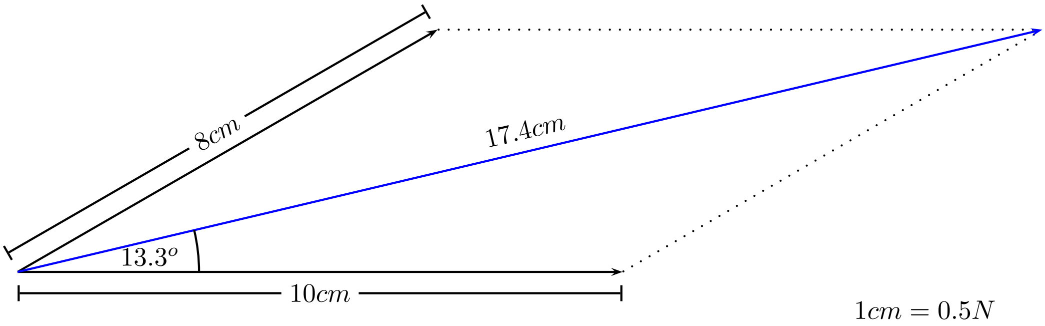
Step 3 : Draw the first scaled vector
Let us draw F_1 first. According to the scale it has length $10cm$:



Step 4 : Draw the second scaled vector
Next we draw F_2. According to the scale it has length $8cm$. We make use of a protractor to draw this vector at 30^o to the horizontal:



Step 5 : Determine the resultant vector
Next we complete the parallelogram and draw the diagonal:



Step 6 : Apply the scale conversion
Finally we use the scale to convert the measured length into the actual magnitude. Since $1cm = 0.5N$, $17.4cm$ represents $8.7N$. Therefore the resultant force is $8.7N$ at 13.3^o above the horizontal.

The parallelogram method is restricted to the addition of just two vectors. However, it is arguably the most intuitive way of adding two forces acting at a point.

4.6.2 Algebraic Addition and Subtraction of Vectors

Vectors in a Straight Line

Whenever you are faced with adding vectors acting in a straight line (i.e. some directed left and some right, or some acting up and others down) you can use a very simple algebraic technique:

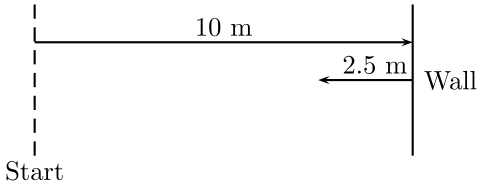
- Choose a positive direction. As an example, for situations involving displacements in the directions west and east, you might choose west as your positive direction. In that case, displacements east are negative.
- Next simply add (or subtract) the vectors with the appropriate signs.
- As a final step the direction of the resultant should be included in words (positive answers are in the positive direction, while negative resultants are in the negative direction).

Let us consider a couple of examples.

Worked Example 8

Adding vectors algebraically I

Question: A tennis ball is rolled towards a wall which is $10m$ away to the right. If after striking the wall the ball rolls a further $2.5m$ along the ground to the left, calculate algebraically the ball's resultant displacement.
(NOTE TO SELF: PGCE suggest a 'more real looking' diagram, followed by a diagram one would draw to solve the problem (like our existing one with the positive direction shown as an arrow)) **Answer:**
Step 1 : Draw a rough sketch of the situation



Step 2 : Decide which method to use to calculate the resultant
We know that the resultant displacement of the ball ($\overrightarrow{s}_{resultant}$) is equal to the sum of the ball's separate displacements ($\overrightarrow{s}_1$ and $\overrightarrow{s}_2$):

$$\overrightarrow{s}_{resultant} = \overrightarrow{s}_1 + \overrightarrow{s}_2$$

Since the motion of the ball is in a straight line (i.e. the ball moves left and right), we can use the method of algebraic addition just explained.

Step 3 : Choose a positive direction
Let's make to the right the **positive** direction. This means that to the left becomes the **negative** direction.
Step 4 : Now define our vectors algebraically
With right positive:

$$\begin{aligned} \overrightarrow{s}_1 &= +10.0m \\ &\textit{and} \\ \overrightarrow{s}_2 &= -2.5m \end{aligned}$$

Step 5 : Add the vectors
Next we simply add the two displacements to give the resultant:

$$\begin{aligned} \overrightarrow{s}_{resultant} &= (+10m) + (-2.5m) \\ &= (+7.5)m \end{aligned}$$

Step 6 : Quote the resultant
Finally, in this case right means positive so:

$$\overrightarrow{s}_{resultant} = 7.5m \text{ to the right}$$

Let us consider an example of vector subtraction.

Worked Example 9

Subtracting vectors algebraically I

Question: Suppose that a tennis ball is thrown horizontally towards a wall at $3m.s^{-1}$ to the right. After striking the wall, the ball returns to the thrower at $2m.s^{-1}$. Determine the change in velocity of the ball.
Answer:
Step 1 : Draw a sketch
A quick sketch will help us understand the problem (NOTE TO SELF: Maybe a sketch here?)
Step 2 : Decide which method to use to calculate the resultant
Remember that velocity is a vector. The change in the velocity of the ball is equal to the difference between the ball's initial and final velocities:

$$\Delta\overrightarrow{v} = \overrightarrow{v}_{final} - \overrightarrow{v}_{initial}$$

Since the ball moves along a straight line (i.e. left and right), we can use the algebraic technique of vector subtraction just discussed.
Step 3 : Choose a positive direction
Let's make to the right the **positive** direction. This means that to the left becomes the **negative** direction.

Step 4 : Now define our vectors algebraically
With right positive:

$$\begin{aligned}\overrightarrow{v}_{initial} &= +3m.s^{-1} \\ &\textit{and} \\ \overrightarrow{v}_{final} &= -2m.s^{-1}\end{aligned}$$

Step 5 : Subtract the vectors
Thus, the change in velocity of the ball is:

$$\begin{aligned}\Delta\overrightarrow{v} &= (-2m.s^{-1}) - (+3m.s^{-1}) \\ &= (-5)m.s^{-1}\end{aligned}$$

Step 6 : Quote the resultant
Remember that in this case right means positive so:

$$\Delta\overrightarrow{v} = 5m.s^{-1} \textbf{ to the } \textit{left}$$

Remember that the technique of addition and subtraction just discussed can only be applied to vectors acting along a straight line.

A More General Algebraic technique

In worked example 3 the tail to head method of accurate construction was used to determine the resultant displacement of a man who travelled first east and then north. However, the man's resultant can be calculated without drawing an accurate scale diagram. Let us revisit this example.

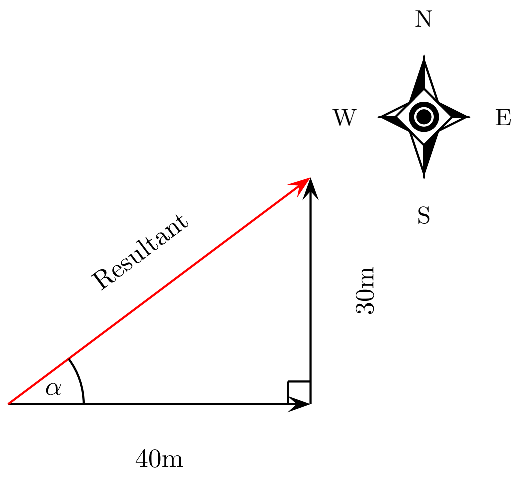
Worked Example 10

An Algebraic solution to Worked Example 3

Question: A man walks 40 m East, then 30 m North.
Calculate the man's resultant displacement.

Answer:

Step 1 : Draw a rough sketch
As before, the rough sketch looks as follows:



Step 2 : Determine the length of the resultant
Note that the triangle formed by his separate displacement vectors and his resultant displacement vector is a right-angle triangle. We can thus use Pythogoras' theorem to determine the length of the resultant. If the length of the resultant vector is called s then:

$$\begin{aligned} s^2 &= (40m)^2 + (30m)^2 \\ s^2 &= 2500m^2 \\ s &= 50m \end{aligned}$$

Step 3 : Determine the direction of the resultant
Now we have the length of the resultant displacement vector but not yet its direction. To determine its direction we calculate the angle α between the resultant displacement vector and East.
We can do this using simple trigonometry:

$$\begin{aligned} \tan\alpha &= \frac{opposite}{adjacent} \\ \tan\alpha &= \frac{30}{40} \\ \alpha &= \arctan(0.75) \\ \alpha &= 36.9^o \end{aligned}$$

Step 4 : Quote the resultant
Our final answer is then:

Resultant Displacement: 50 m at 36.9^o North of East

This is exactly the same answer we arrived at after drawing a scale diagram!

In the previous example we were able to use simple trigonometry to calculate a man's resultant displacement. This was possible since the man's directions of motion were perpendicular (north and east). Algebraic techniques, however, are not limited to cases where the vectors to be combined are along the same straight line or at right angles to one another. The following example illustrates this.

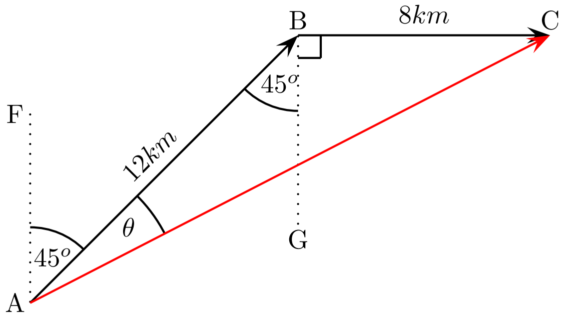
Worked Example 11

Further example of vector addition by calculation

Question: A man walks from point A to point B which is $12km$ away on a bearing of 45^o. From point B the man walks a further $8km$ east to point C. Calculate the man's resultant displacement.

Answer:

Step 1 : Draw a rough sketch of the situation



$B\hat{A}F = 45^o$ since the man walks initially on a bearing of 45^o. Then, $A\hat{B}G = B\hat{A}F = 45^o$ (alternate angles parallel lines). Both of these angles are included in the rough sketch.

Step 2 : Calculate the length of the resultant

The resultant is the vector AC. Since we know both the lengths of AB and BC and the included angle $A\hat{B}C$, we can use the cosine rule:

$$\begin{aligned} AC^2 &= AB^2 + BC^2 - 2 \cdot AB \cdot BC \cos(A\hat{B}C) \\ &= (12)^2 + (8)^2 - 2 \cdot (12)(8) \cos(135^o) \\ &= 343.8 \\ AC &= 18.5\ km \end{aligned}$$

Step 3 : Determine the direction of the resultant

Next we use the sine rule to determine the angle θ:

$$
\begin{aligned}
\frac{\sin\theta}{8} &= \frac{\sin 135^0}{18.5} \\
\sin\theta &= \frac{8 \times \sin 135^o}{18.5} \\
\theta &= \arcsin(0.3058) \\
\theta &= 17.8^o
\end{aligned}
$$

Thus, $F\hat{A}C = 62.8^o$.

Step 4 : Quote the resultant

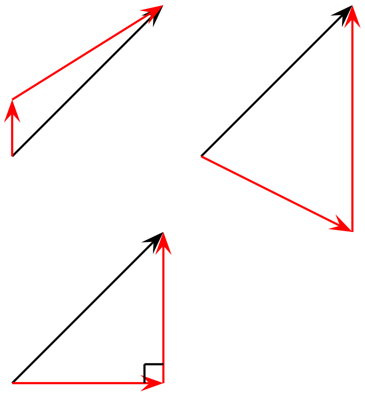
Our final answer is then:

Resultant Displacement: $18.5km$ on a bearing of 62.8^o

4.7 Components of Vectors

In the discussion of vector addition we saw that a number of vectors acting together can be combined to give a single vector (the resultant). In much the same way a single vector can be broken down into a number of vectors which when added give that original vector. These vectors which sum to the original are called **components** of the original vector. The process of breaking a vector into its components is called **resolving into components**.

While summing a given set of vectors gives just one answer (the resultant), a single vector can be resolved into infinitely many sets of components. In the diagrams below the same black vector is resolved into different pairs of components. These components are shown in red. When added together the red vectors give the original black vector (i.e. the original vector is the resultant of its components).



In practice it is most useful to resolve a vector into components which are at right angles to one another.

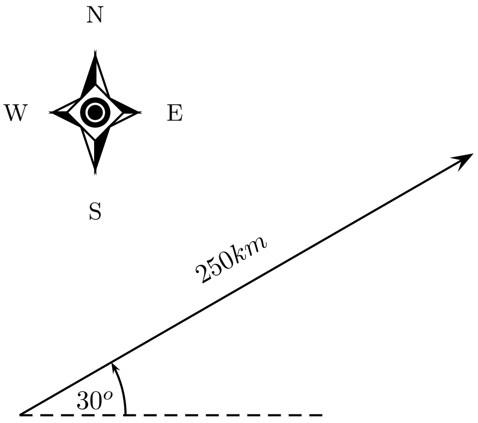
Worked Example 12

Resolving a vector into components

Question: A motorist undergoes a displacement of $250km$ in a direction 30^o north of east. Resolve this displacement into components in the directions north ($\overrightarrow{s}_N$) and east ($\overrightarrow{s}_E$).

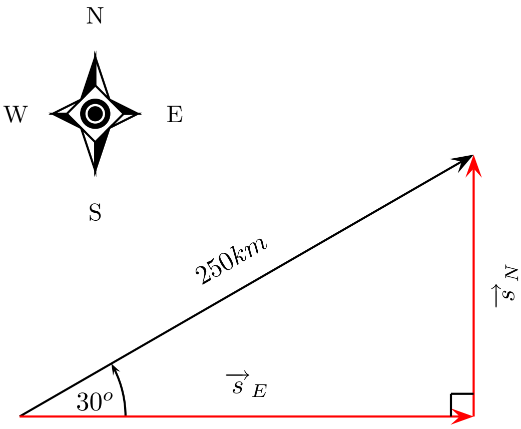
Answer:

Step 1 : Draw a rough sketch of the original vector



Step 2 : Determine the vector component

Next we resolve the displacement into its components north and east. Since these directions are orthogonal to one another, the components form a right-angled triangle with the original displacement as its hypotenuse:



Notice how the two components acting together give the orginal vector as their resultant.

Step 3 : Determine the lengths of the component vectors

Now we can use trigonometry to calculate the magnitudes of the components of the original displacement:

$$\begin{aligned} s_N &= 250 \sin 30^o \\ &= 125\ km \end{aligned}$$

and

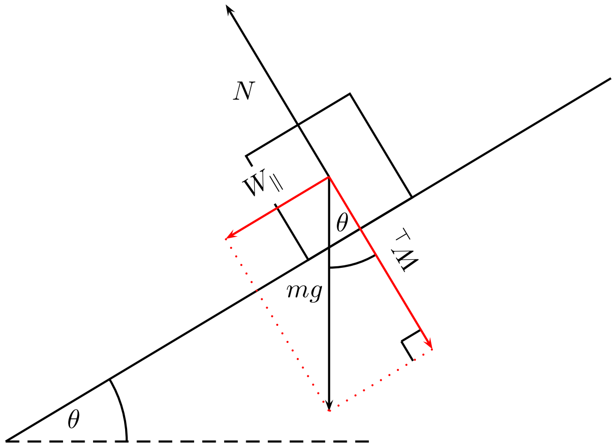
$$\begin{aligned} s_E &= 250\cos 30^o \\ &= 216.5\ km \end{aligned}$$

Remember s_N and s_E are the magnitudes of the components– they are in the directions north and east respectively.
(NOTE TO SELF: SW: alternatively these results can be arrived at by construction. Include?)

4.7.1 Block on an incline

As a further example of components let us consider a block of mass m placed on a frictionless surface inclined at some angle θ to the horizontal. The block will obviously slide down the incline, but what causes this motion?

The forces acting on the block are its weight mg and the normal force N exerted by the surface on the object. These two forces are shown in the diagram below.



Now the object's weight can be resolved into components parallel and perpendicular to the inclined surface. These components are shown as red arrows in the diagram above and are at right angles to each other. The components have been drawn acting from the same point. Applying the parallelogram method, the two components of the block's weight sum to the weight vector.

To find the components in terms of the weight we can use trigonometry:

$$\begin{aligned} W_{||} &= mg\sin\theta \\ W_\perp &= mg\cos\theta \end{aligned}$$

The component of the weight perpendicular to the slope $W_\perp$ exactly balances the normal force N exerted by the surface. The parallel component, however, $W_{||}$ is unbalanced and causes the block to slide down the slope.

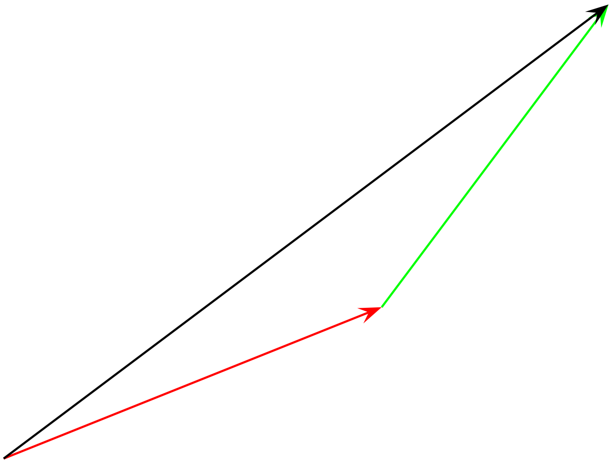


Figure 4.2: An example of two vectors being added to give a resultant

4.7.2 Vector addition using components

In Fig 4.3 two vectors are added in a slightly different way to the methods discussed so far. It might look a little like we are making more work for ourselves, but in the long run things will be easier and we will be less likely to go wrong.

In Fig 4.3 the primary vectors we are adding are represented by solid lines and are the same vectors as those added in Fig 4.2 using the less complicated looking method.

Each vector can be broken down into a component in the x-direction and one in the y-direction. These components are two vectors which when added give you the original vector as the resultant. Look at the red vector in figure 4.3. If you add up the two red dotted ones in the x-direction and y-direction you get the same vector. For all three vectors we have shown their respective components as dotted lines in the same colour.

But if we look carefully, addition of the x components of the two original vectors gives the x component of the resultant. The same applies to the y components. So if we just added all the components together we would get the same answer! This is another important property of vectors.

Worked Example 13

Adding Vectors Using Components

Question: Lets work through the example shown in Fig. 4.3 to determine the resultant.

Answer:

Step 1 : Decide how to tackle the problem

The first thing we must realise is that the order that we add the vectors does not matter. Therefore, we can work through the vectors to be added in any order.

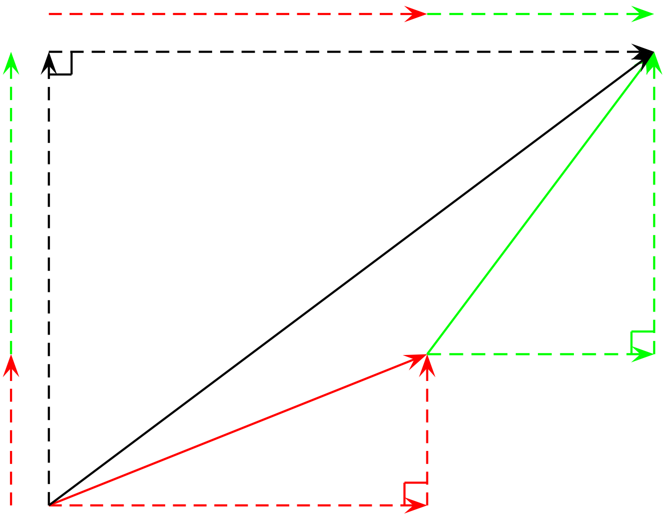
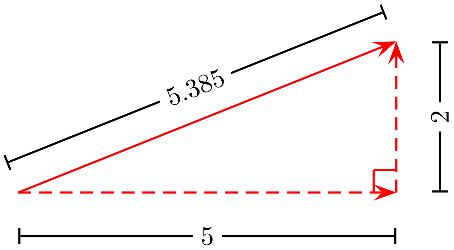


Figure 4.3: Components of vectors can be added as well as the vectors themselves

Step 2 : Resolve the red vector into components

Let us start with the bottom vector. If you are told that this vector has a length of 5.385 units and an angle of 21.8^o to the horizontal then we can find its components. We do this by using known trigonometric ratios. First we find the vertical or y component:

$$\begin{aligned} \sin\theta &= \frac{y}{\text{hypotenuse}} \\ \sin(21.8) &= \frac{y}{5.385} \\ y &= 5.385\sin(21.8) \\ y &= 2 \end{aligned}$$



Secondly we find the horizontal or x component:

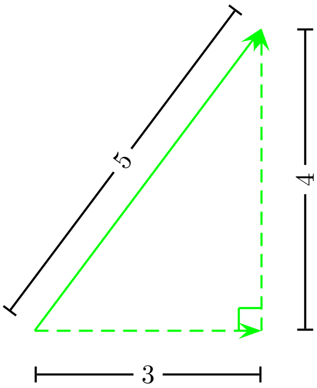
$$\begin{aligned} \cos\theta &= \frac{x}{\text{hypotenuse}} \\ \cos(21.8) &= \frac{x}{5.385} \\ x &= 5.385\cos(21.8) \\ x &= 5 \end{aligned}$$

We now know the lengths of the sides of the triangle for which our vector is the hypotenuse. If you look at these sides we can assign them directions given by the dotted arrows. Then our original red vector is just the sum of the two dotted vectors (its components). When we try to find the final answer we can just add all the dotted vectors because they would add up to the two vectors we want to add.

Step 3 : Now resolve the second vector into components

. The green vector has a length of 5 units and a direction of 53.13 degrees to the horizontal so we can find its components.

$$\begin{aligned} \sin\theta &= \frac{y}{\text{hypotenuse}} \\ \sin(53.13) &= \frac{y}{5} \\ y &= 5\sin(53.13) \\ y &= 4 \end{aligned}$$



$$\begin{aligned} \cos\theta &= \frac{x}{\text{hypotenuse}} \\ \cos(53.13) &= \frac{x}{5} \\ x &= 5\cos(53.13) \\ x &= 3 \end{aligned}$$

Step 4 : Determine the components of the resultant vector

Now we have all the components. If we add all the x-components then we will have the x-component of the resultant vector. Similarly if we add all the y-components then we will have the y-component of the resultant vector.

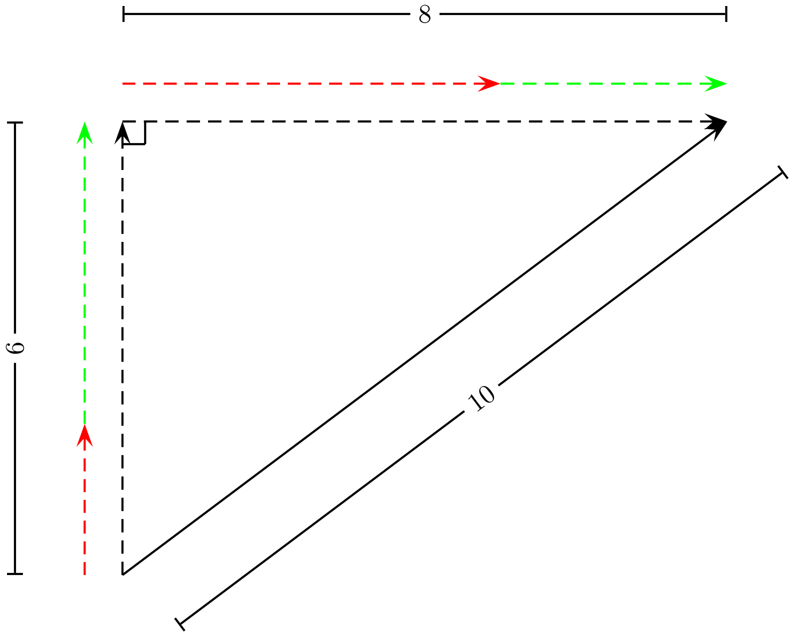
The x-components of the two vectors are 5 units right and then 3 units right. This gives us a final x-component of 8 units right.

The y-components of the two vectors are 2 units up and then 4 units up. This gives us a final y-component of 6 units up.

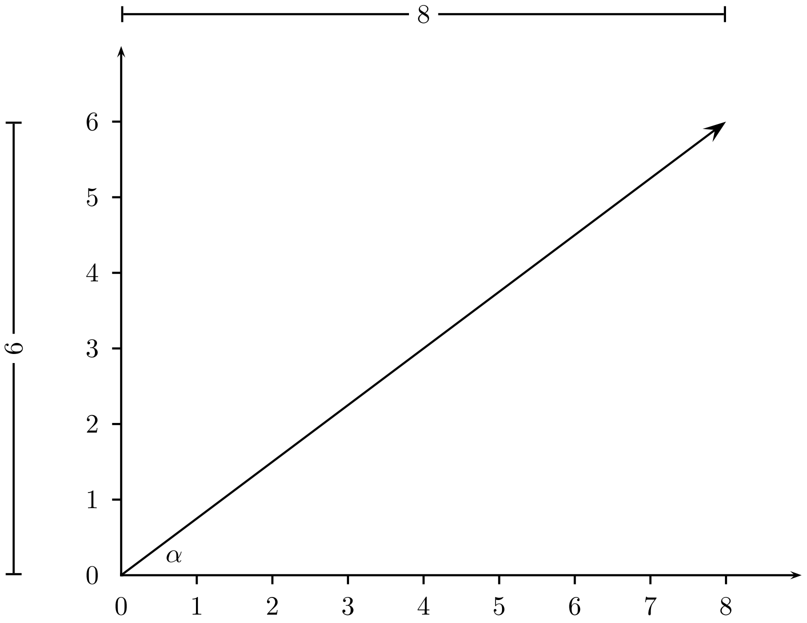
Step 5 : Determine the magnitude and direction of the resultant vector

Now that we have the components of the resultant, we can use Pythagoras' theorem to determine the length of the resultant. Let us call the length of the hypotenuse l and we can calculate its value;

$$
\begin{aligned}
l^2 &= (6)^2 + (8)^2 \\
l^2 &= 100 \\
l &= 10.
\end{aligned}
$$



The resultant has length of 10 units so all we have to do is calculate its direction. We can specify the direction as the angle the vectors makes with a known direction. To do this you only need to visualise the vector as starting at the origin of a coordinate system. We have drawn this explicity below and the angle we will calculate is labeled α.



Using our known trigonometric ratios we can calculate the value of α;

$$
\begin{aligned}
\tan\alpha &= \frac{6}{8} \\
\alpha &= \arctan\frac{6}{8} \\
\alpha &= 36.8^o.
\end{aligned}
$$

Step 6 : Quote the final answer
Our final answer is a resultant of 10 units at 36.8^o to the positive x-axis.

4.8 Do I really need to learn about vectors? Are they really useful?

Vectors are essential to do physics. Absolutely essential. This is an important warning. If something is essential we had better stop for a moment and make sure we understand it properly.

4.9 Summary of Important Quantities, Equations and Concepts

Vector A *vector* is a measurement which has both magnitude and direction.

Displacement *Displacement* is a vector with direction pointing from some initial (starting) point to some final (end) point and whose magnitude is the straight-line distance from the starting point to the final point.

Distance The *distance* travelled is the length of your actual path.

Velocity *Velocity* is the rate of change of displacement with respect to time.

Acceleration *Acceleration* is the rate of change of velocity with respect to time.

Resultant The *resultant* of a number of vectors is the single vector whose effect is the same as the individual vectors acting together.

Quantity	Symbol	S.I. Units	Direction
Displacement	$\overrightarrow{s}$	m	✓
Velocity	$\overrightarrow{u}, \overrightarrow{v}$	$m.s^{-1}$	✓
Distance	d	m	–
Speed	v	$m.s^{-1}$	–
Acceleration	$\overrightarrow{a}$	$m.s^{-2}$	✓

Table 4.1: Summary of the symbols and units of the quantities used in **Vectors**

Chapter 5

Forces

5.1 'TO DO' LIST

- introduce concept of a system for use in momentum- in NIII for instance, also concept of isolated system and external forces
- add incline plane examples
- generally 'bulk-up' the Newton's Laws sections

5.2 What is a force?

The simplest answer is to say a 'push' or a 'pull'. If the force is great enough to overcome friction the object being pushed or pulled will move. We could say a force is something that makes objects move. Actually forces give rise to accelerations!

In fact, the acceleration of a body is directly proportional to the *net* force acting on it. The word *net* is important– forces are vectors and what matters in any situation is the *vector sum* of all the forces acting on an object.

The unit of force is the *newton* (symbol N). It is named after Sir Isaac Newton, whose three laws you will learn about shortly.

Interesting Fact: Force was first described by Archimedes. Archimedes of Syracuse (circa 287 BC - 212 BC), was a Greek mathematician, astronomer, philosopher, physicist and engineer. He was killed by a Roman soldier during the sack of the city, despite orders from the Roman general, Marcellus, that he was not to be harmed.

5.3 Force diagrams

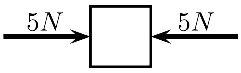
The *resultant force* acting on an object is the vector sum of the set of forces acting on that one object. It is very important to remember that all the forces must be acting on the *same* object.

The easiest way to determine this resultant force is to construct what we call a force diagram. In a force diagram we represent the object by a point and draw all the force vectors connected to that point as arrows. Remember from Chapter **??** that we use the length of the arrow to indicate the vector's magnitude and the direction of the arrow to show which direction it acts in.

The second step is to rearrange the force vectors so that it is easy to add them together and find the resultant force.

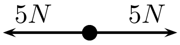
Let us consider an example to get started:

Two people push on a box from opposite sides with a force of $5N$.



When we draw the force diagram we represent the box by a dot. The two forces are represented by arrows, with their tails on the dot.

Force Diagram:



See how the arrows point in opposite directions and have the same magnitude (length). This means that they cancel out and there is no *net* force acting on the object.

This result can be obtained algebraically too, since the two forces act along the same line. Firstly we choose a positive direction and then add the two vectors taking their directions into account.

Consider to the right as the positive direction

$$\begin{aligned} F_{res} &= (+5N) + (-5N) \\ &= 0N \end{aligned}$$

Always rememb
to check you
signs

As you work with more complex force diagrams, in which the forces do not exactly balance, you may notice that sometimes you get a negative answer (e.g. $-2N$). What does this mean? Does it mean that we have something the opposite of force? No, all it means is that the force acts in the *opposite* direction to the one that you chose to be positive. You can *choose* the positive direction to be any way you want, but once you have chosen it you *must* keep it.

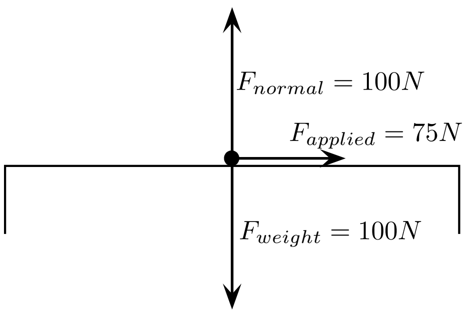
Once a force diagram has been drawn the techniques of vector addition introduced in the previous chapter can be implemented. Depending on the situation you might choose to use a graphical technique such as the tail-to-head method or the parallelogram method, or else an algebraic approach to determine the resultant. Since force is a vector all of these methods apply!

Worked Example 14

Single Force on a block

Question: A block on a frictionless flat surface weighs $100N$. A $75N$ force is applied to the block towards the right. What is the net force (or resultant force) on the block?
Answer:

Step 1 : Firstly, draw a force diagram for the block



Be careful not to forget the two forces perpendicular to the surface. Every object with mass is attracted to the centre of the earth with a force (the object's weight). However, if this were the only force acting on the block in the vertical direction then the block would fall through the table to the ground. This does not happen because the table exerts an upward force (the normal force) which exactly balances the object's weight.

Step 2 : Answer

Thus, the only unbalanced force is the applied force. This applied force is then the resultant force acting on the block.

5.4 Equilibrium of Forces

At the beginning of this chapter it was mentioned that resultant forces cause objects to accelerate. If an object is stationary or moving at constant velocity then either,

- no forces are acting on the object, or
- the forces acting on that object are exactly balanced.

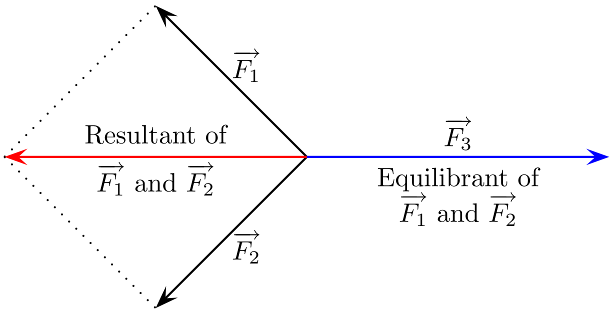
A resultant force would cause a stationary object to start moving or an object moving at constant velocity to speed up or slow down.

In other words, for stationary objects or objects moving with constant velocity, the resultant force acting on the object is zero. The object is said to be in **equilibrium**.

Objects at re
or moving wi
constant veloc
are in **equilibr**
and have a **ze**
resultant fo

If a resultant force acts on an object then that object can be brought into equilibrium by applying an additional force that exactly balances this resultant. Such a force is called the *equilibrant* and is equal in magnitude but opposite in direction to the original resultant force acting on the object.

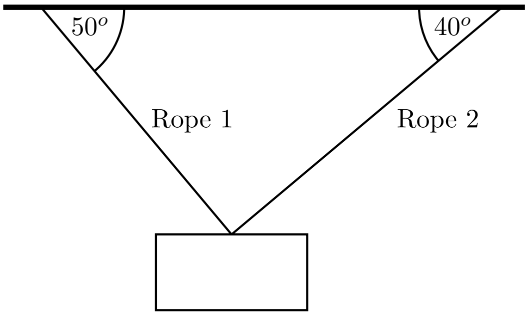
Definition: The *equilibrant* of any number of forces is the single force required to produce equilibrium.


$\overrightarrow{F_1}$
Resultant of
$\overrightarrow{F_1}$ and $\overrightarrow{F_2}$
$\overrightarrow{F_3}$
Equilibrant of
$\overrightarrow{F_1}$ and $\overrightarrow{F_2}$
$\overrightarrow{F_2}$

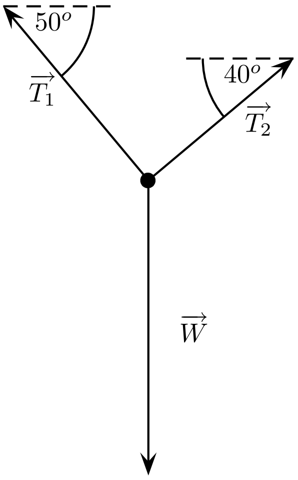
In the figure the resultant of $\overrightarrow{F_1}$ and $\overrightarrow{F_2}$ is shown in red. The equilibrant of $\overrightarrow{F_1}$ and $\overrightarrow{F_2}$ is then the vector opposite in direction to this resultant with the same magnitude (i.e. $\overrightarrow{F_3}$).

- $\overrightarrow{F_1}$, $\overrightarrow{F_2}$ and $\overrightarrow{F_3}$ are in equilibrium
- $\overrightarrow{F_3}$ is the equilibrant of $\overrightarrow{F_1}$ and $\overrightarrow{F_2}$
- $\overrightarrow{F_1}$ and $\overrightarrow{F_2}$ are kept in equilibrium by $\overrightarrow{F_3}$

As an example of an object in equilibrium, consider an object held stationary by two ropes in the arrangement below:

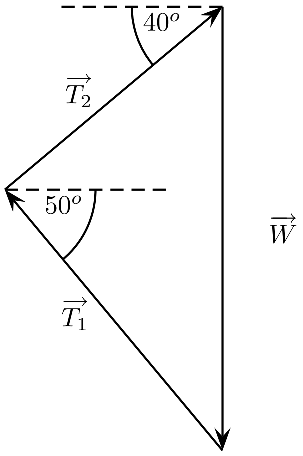


Let us draw a force diagram for the object. In the force diagram the object is drawn as a dot and all forces acting on the object are drawn in the correct directions starting from that dot. In this case, three forces are acting on the object.



Each rope exerts a force on the object in the direction of the rope away from the object. These tension forces are represented by $\overrightarrow{T_1}$ and $\overrightarrow{T_2}$. Since the object has mass, it is attracted towards the centre of the earth. This weight is represented in the force diagram as $\overrightarrow{W}$.

Since the object is stationary, the resultant force acting on the object is zero. In other words the three force vectors drawn tail-to-head form a closed triangle:



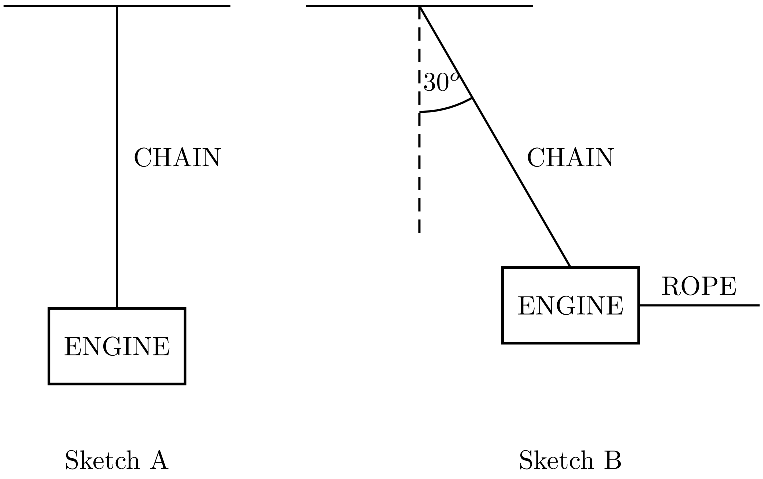
In general, when drawn tail-to-head the forces acting on an object in equilibrium form a closed figure with the head of the last vector joining up with the tail of the first vector. When only three forces act on an object this closed figure is a triangle. This leads to the **triangle law for three forces in equilibrium**:

Triangle Law for Three Forces in Equilibrium:
Three forces in equilibrium can be represented in magnitude and direction by the three sides of a triangle taken in order.

Worked Example 15

Equilibrium

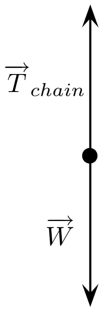
Question: A car engine of weight $2000N$ is lifted by means of a chain and pulley system. In sketch A below, the engine is suspended by the chain, hanging stationary. In sketch B, the engine is pulled sideways by a mechanic, using a rope. The engine is held in such a position that the chain makes an angle of 30^o with the vertical. In the questions that follow, the masses of the chain and the rope can be ignored.



i) Draw a force diagram representing the forces acting on the engine in sketch A.
ii) Determine the tension in the chain in sketch A.
iii) Draw a force diagram representing the forces acting on the engine in sketch B.
ii) In sketch B determine the magnitude of the applied force and the tension in the chain.

Answer:

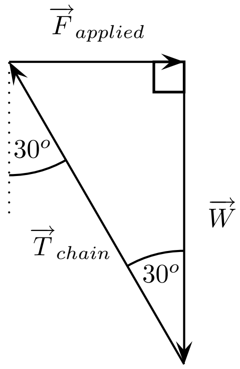
Step 1 : Force diagram for sketch A
i) Just two forces are acting on the engine in sketch A:



Step 2 : Determine the tension in the chain
ii) Since the engine in sketch A is stationary, the resultant force on the engine is zero. Thus the tension in the chain exactly balances the weight of the engine,

$$\begin{aligned} T_{chain} &= W \\ &= 2000N \end{aligned}$$

Step 3 : Force diagram for sketch B
iii) Three forces are acting on the engine in sketch B:



Since the engine is at equilibrium (it is held stationary) the three forces drawn tail-to-head form a closed triangle.
Step 4 : Calculate the magnitude of the forces in sketch B
iv) Since no method was specified let us calculate the magnitudes algebraically. Since the triangle formed by the three forces is a right-angle triangle this is easily done:

$$\begin{aligned} \frac{F_{applied}}{W} &= \tan 30^o \\ F_{applied} &= (2000)\tan 30^o \\ &= 1155\ N \end{aligned}$$

and

$$\begin{aligned} \frac{T_{chain}}{W} &= \frac{1}{\cos 30^o} \\ T_{chain} &= \frac{2000}{\cos 30^o} \\ &= 2309\ N \end{aligned}$$

5.5 Newton's Laws of Motion

Our current laws of motion were discovered by Sir Isaac Newton. It is said that Sir Isaac Newton started to think about the nature of motion and gravitation after being struck on the head by a falling apple.

Newton discovered 3 laws describing motion:

5.5.1 First Law

> **Definition:** Every object will remain at rest or in uniform motion in a straight line unless it is made to change its state by the action of an *external force*.

For example, a cement block isn't going to move unless you push it. A rocket in space is not going to speed up, slow down nor change direction unless the engines are switched on.

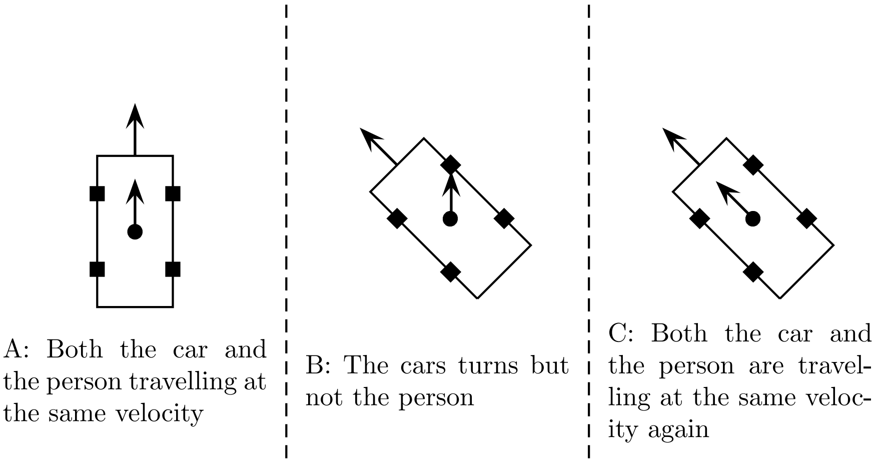
Newton's First Law may seem rather surprising to you when you first meet it. If you roll a ball along a surface it always stops, but Newton's First Law says that *an item will remain in uniform motion in a straight line unless a force acts upon it.* It doesn't seem like there is any force acting on the ball once you let it go, but there is! In real life (unlike many physics problems) there is usually friction acting. Friction is the *external* force acting on the ball causing it to stop.

Notice that Newton's First Law says that the force must be *external.* For example, you can't grab your belt and pull yourself up to the ceiling. Of course, you could get someone else to pull you up, but then that person would be applying an *external* force.

Worked Example 16

Newton's First Law in action

Question: Why do passengers get thrown to the side when the car they are driving in goes around a corner?
Answer: Newton's First Law
Step 1 : What happens before the car turns
Before the car starts turning both you and the car are travelling at the same velocity. (picture A)
Step 2 : What happens while the car turns
The driver turns the wheels of the car, which then exert a force on the car and the car turns. This force acts on the car but not you, hence (by Newton's First Law) you continue moving with the same velocity. (picture B)
Step 3 : Why passengers get thrown to the side
If the passenger is wearing a seatbelt it will exert a force on the passenger until the passenger's velocity is the same as that of the car (picture C). Without a seatbelt the passenger may hit the side of the car or even the windscreen!



A: Both the car and the person travelling at the same velocity

B: The cars turns but not the person

C: Both the car and the person are travelling at the same velocity again

5.5.2 Second Law

> **Definition:** The resultant force acting on a body results in an acceleration which is in the same direction as the resultant force and is directly proportional to the magnitude of this force and inversely proportional to the mass of the object.

In mathematical form this law states,

$$\overrightarrow{F}_{Res} = m\overrightarrow{a} \tag{5.1}$$

So if a resultant force $\overrightarrow{F}_{Res}$ acts on an object with mass m it will result in an acceleration $\overrightarrow{a}$. It makes sense that the direction of the acceleration is in the direction of the resultant force. If you push something away from you it doesn't move toward you unless of course there is another force acting on the object towards you!

Worked Example 17

Newton's Second Law

Question: A block of mass $10kg$ is accelerating at $2m.s^{-2}$. What is the magnitude of the resultant force acting on the block?
Answer:
Step 1 : Decide what information has been supplied
We are given

- the block's mass
- the block's acceleration

all in the correct units.

Step 2 : Determine the force on the block
We are asked to find the magnitude of the force applied to the block. Newton's Second Law tells us the relationship between acceleration and force for an object. Since we are only asked for the magnitude we do not need to worry about the directions of the vectors:

$$\begin{aligned} F_{Res} &= ma \\ &= 10kg \times 2m.s^{-2} \\ &= 20N \end{aligned}$$

Thus, there must be a resultant force of $20N$ acting on the box.

Worked Example 18

Newton's Second Law 2

Question: A $12N$ force is applied in the positive x-direction to a block of mass $100mg$ resting on a frictionless flat surface. What is the resulting acceleration of the block?
Answer:
Step 1 : Decide what information has been supplied
We are given

- the block's mass
- the applied force

but the mass is not in the correct units.
Step 2 : Convert the mass into the correct units

$$\begin{aligned} 100mg &= 100 \times 10^{-3}g = 0.1g \\ 1000g &= 1kg \\ 1 &= 1kg \times \frac{1}{1000g} \\ &= \frac{1kg}{1000g} \\ 0.1g &= 0.1g \times 1 \\ &= 0.1g \times \frac{1kg}{1000g} \\ &= 0.0001\ kg \end{aligned}$$

Step 3 : Determine the direction of the acceleration
We know that net force results in acceleration. Since there is no friction the applied force is the resultant or net force on the block (refer to the earlier example of the block pushed on the surface of the table). The block will then accelerate in the direction of this force according to Newton's Second Law.

Step 4 : Determine the magnitude of the acceleration

$$\begin{aligned}
F_{Res} &= ma \\
12N &= (0.0001kg)a \\
a &= \frac{12N}{0.0001kg} \\
&= 120000\frac{N}{kg} \\
&= 120000\frac{kg.m}{s^2.kg} \\
&= 120000\frac{m}{s^2} \\
&= 1.2 \times 10^5\ m.s^{-2}
\end{aligned}$$

From Newton's Second Law the direction of the acceleration is the same as that of the resultant force. The final result is then that the block accelerates at $1.2 \times 10^5\ m.s^{-2}$ in the positive x-direction.

Weight and Mass

You must have heard people saying "My *weight* is 60kg". This is actually incorrect because it is *mass* that is measured in kilograms. *Weight* is the force of gravity exerted by the earth on an object with mass:

$$F_{weight} = mg \tag{5.2}$$

As such, weight is measured in newtons.

If you compare this equation to Newton's Second Law you will see that it looks exactly the same with the a replaced by g. Thus, when weight is the only force acting on an object (i.e. when F_{weight} is the resultant force acting on the object) the object has an acceleration g. Such an object is said to be in free fall. The value for g is the same for all objects (i.e. it is independent of the objects mass):

$$g = 9.8ms^{-2} \approx 10ms^{-2} \tag{5.3}$$

You will learn how to calculate this value from the mass and radius of the earth in Chapter **??**. Actually the value of g varies slightly from place to place on the earth's surface.

The reason that we often get confused between weight and mass, is that scales measure your weight (in newtons) and then display your mass using the equation above.

(NOTE TO SELF: include example of skydiver- calculate the acceleration before and after opening parachute– free-fall and then negative acceleration)

Worked Example 19

Calculating the resultant and then the acceleration

Question: A block (mass $20kg$) on a frictionless flat surface has a $45N$ force applied to it in the positive x-direction. In addition a $25N$ force is applied in the negative x-direction. What is the resultant force acting on the block and the acceleration of the block?
Answer:
Step 1 : Decide what information has been supplied
We are given

- the block's mass
- Force $F_1 = 45N$ in the positive x-direction
- Force $F_2 = 25N$ in the negative x-direction

all in the correct units.
Step 2 : Figure out how to tackle the problem
We are asked to determine what happens to the block. We know that net force results in an acceleration. We need to determine the net force acting on the block.
Step 3 : Determine the magnitude and direction of the resultant force
Since F_1 and F_2 act along the same straight line, we can apply the algebraic technique of vector addition discussed in the Vectors chapter to determine the resultant force. Choosing the positive x-direction as our positive direction:

Positive x-direction is the positive direction:

$$\begin{aligned} F_{Res} &= (+45N) + (-25N) \\ &= +20N \\ &= 20N\ in\ the\ positive\ x-direction \end{aligned}$$

where we remembered in the last step to include the direction of the resultant force in words. By Newton's Second Law the block will accelerate in the direction of this resultant force.
Step 4 : Determine the magnitude of the acceleration

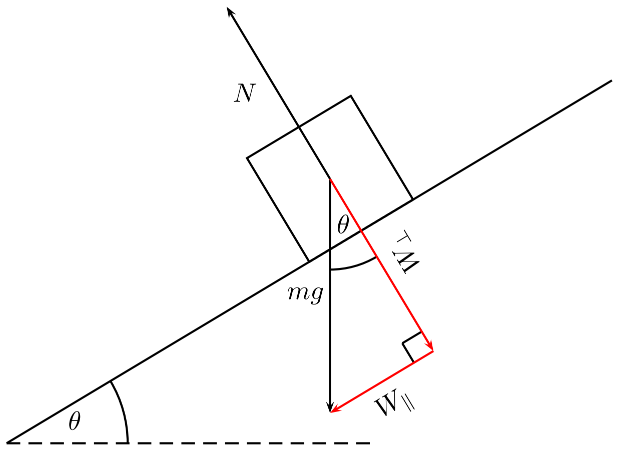
$$\begin{aligned} F_{Res} &= ma \\ 20N &= (20kg)a \\ a &= \frac{20N}{20kg} \\ &= 1\frac{N}{kg} \\ &= 1\frac{kg.m}{s^2.kg} \\ &= 1\ m.s^{-2} \end{aligned}$$

The final result is then that the block accelerates at 1 $m.s^{-2}$ in the positive x-direction (the same direction as the resultant force).

Worked Example 20

Block on incline

Question: insert question here maybe with friction!



5.5.3 Third Law

Definition: For every force or action there is an *equal* but *opposite* force or reaction.

Newton's Third Law is easy to understand but it can get quite difficult to apply it. An important thing to realise is that the action and reaction forces can never act on the same object and hence cannot contribute to the same resultant.

Worked Example 21

Identifying action-reaction pairs

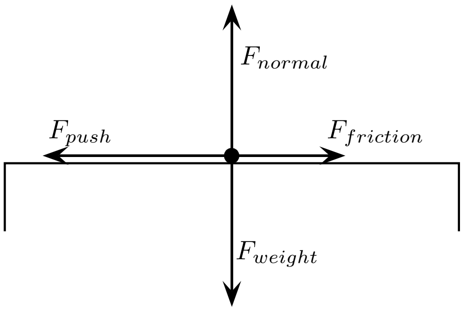
Question: Consider pushing a box on the surface of a rough table.

1. Draw a force diagram indicating all of the forces acting **on the box**.
2. Identify the reaction force for each of the forces acting on the box.

Answer:

1. The following force diagram shows all of the forces acting *on the box*

Remember to d
the object in y
force diagram
dot



2. The following table lists each of forces acting on the box (the actions) together with their corresponding reactions:

Action	**Reaction**
F_{push} : Person pushing on the box	Box pushing on the person
F_{weight} : The earth attracting the box (the weight)	The box attracting the earth
F_{normal} : The box pushing on the table	The table pushing on the box
$F_{friction}$: The table acting on the box	The box acting on the table

Notice that to find the reaction force you need to switch around the object supplying the force and the object receiving the force. Be careful not to think that the normal force is the reaction partner to the weight of the box. The normal force balances the weight force but they are both forces acting on the box so they cannot possibly form an action-reaction pair.

There is an important thing to realise which is related to Newton's Third Law. Think about dropping a stone off a cliff. It falls because the earth exerts a force on it (see Chapter **??**) and it doesn't seem like there are any other forces acting. So is Newton's Third Law wrong? No, the reactionary force to the weight of the stone is the force exerted by the stone on the earth. This is illustrated in detail in the next worked example.

Worked Example 22

Newton's Third Law

Question: A stone of mass $0.5kg$ is accelerating at 10 $m.s^{-2}$ towards the earth.

1. What is the force exerted by the earth on the stone?
2. What is the force exerted by the stone on the earth?
3. What is the acceleration of the earth, given that its mass is $5.97 \times 10^{27} kg$?

Answer:

1. *Step 1 : Decide what information has been supplied*

   We are given

   - the stone's mass
   - the stone's acceleration (g)

   all in the correct units.

   Step 2 : The force applied by the earth on the stone

   This force is simply the weight of the stone.

   Step 3 : The magnitude of the force of the earth on the stone

   Applying Newton's Second Law, we can find the magnitude of this force,

$$\begin{aligned} F_{Res} &= ma \\ &= 0.5kg \times 10\ m.s^{-2} \\ &= 5N \end{aligned}$$

   Therefore the earth applies a force of $5N$ towards the earth on the stone.

2. By Newton's Third Law the stone must exert an equal but opposite force on the earth. Hence the stone exerts a force of $5N$ towards the stone on the earth.

3. We have

   - the force acting on the earth
   - the earth's mass

   in the correct units

   Step 4 : To find the earth's acceleration we must apply Newton's Second Law to find the magnitude

$$\begin{aligned} F_{Res} &= ma \\ 5N &= (5.97 \times 10^{27} kg)a \\ a &= \frac{5N}{5.97 \times 10^{27} kg} \\ &= 8.37521 \times 10^{-28}\ m.s^{-2} \end{aligned}$$

   The earth's acceleration is directed towards the stone (i.e. in the same direction as the force on the earth). The earth's acceleration is really tiny. This is why you don't notice the earth moving towards the stone, even though it does.

Interesting Fact: Newton first published these laws in *Philosophiae Naturalis Principia Mathematica (1687)* and used them to prove many results concerning the motion of physical objects. Only in 1916 were Newton's Laws superceded by Einstein's theory of relativity.

The next two worked examples are quite long and involved but it is very important that you understand the discussion as they illustrate the importance of Newton's Laws.

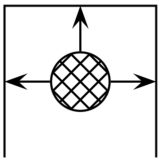
Worked Example 23

Rockets

Question: How do rockets accelerate in space?

Answer:

- Gas explodes inside the rocket.
- This exploding gas exerts a force on each side of the rocket (as shown in the picture below of the explosion chamber inside the rocket).

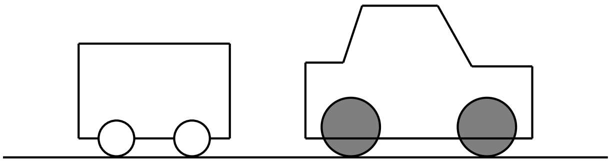


Note that the forces shown in this picture are representative. With an explosion there will be forces in all directions.

- Due to the symmetry of the situation, all the forces exerted on the rocket are balanced by forces on the opposite side, except for the force opposite the open side. This force on the upper surface is unbalanced.
- This is therefore the resultant force acting on the rocket and it makes the rocket accelerate forwards.

Systems and External Forces

The concepts of a system and an external forces are very important in physics. A system is any collection of objects. If one draws an imaginary box around such a system then an external force is one that is applied by an object or person outside the box. Imagine for example a car pulling a trailer.



5.6 Examples of Forces Studied Later

Most of physics revolves around forces. Although there are many different forces we deal with them all in the same way. The methods to find resultants and acceleration do not depend on the type of force we are considering.

At first glance, the number of different forces may seem overwhelming - gravity, drag, electrical forces, friction and many others. However, physicists have found that all these forces can be classified into four groups. These are gravitational forces, electromagnetic forces, strong nuclear force and weak nuclear force. Even better, all the forces that you will come across at school are either gravitational or electromagnetic. Doesn't that make life easy?

5.6.1 Newtonian Gravity

Gravity is the attractive force between two objects due to the mass of the objects. When you throw a ball in the air, its mass and the earth's mass attract each other, which leads to a force between them. The ball falls back towards the earth, and the earth accelerates towards the ball. The movement of the earth toward the ball is, however, so small that you couldn't possibly measure it.

5.6.2 Electromagnetic Forces

Almost all of the forces that we experience in everyday life are electromagnetic in origin. They have this unusual name because long ago people thought that electric forces and magnetic forces were different things. After much work and experimentation, it has been realised that they are actually different manifestations of the same underlying theory.

The Electric Force

If we have objects carrying electrical charge, which are not moving, then we are dealing with electrostatic forces (Coulomb's Law). This force is actually much stronger than gravity. This may seem strange, since gravity is obviously very powerful, and holding a balloon to the wall seems to be the most impressive thing electrostatic forces have done, but think about it: for gravity to be detectable, we need to have a very large mass nearby. But a balloon rubbed in someone's hair can stick to a wall with a force so strong that it overcomes the force of gravity—with just the charges in the balloon and the wall!

Magnetic force

The magnetic force is a different manifestation of the electromagnetic force. It stems from the interaction between *moving charges* as opposed to the *fixed* charges involved in Coulomb's Law.

Examples of the magnetic force in action include magnets, compasses, car engines, computer data storage and your hair standing on end. Magnets are also used in the wrecking industry to pick up cars and move them around sites.

Friction

We all know that Newton's First Law states that an object moving *without a force acting on it* will keep moving. Then why does a box sliding on a table stop? The answer is friction. Friction arises from the interaction between the molecules on the bottom of a box with the molecules on a table. This interaction is electromagnetic in origin, hence friction is just another view of the electromagnetic force. The great part about school physics is that most of the time we are told to neglect friction but it is good to be aware that there is friction in the real world.

Friction is also useful sometimes. If there was no friction and you tried to prop a ladder up against a wall, it would simply slide to the ground. Rock climbers use friction to maintain their grip on cliffs.

Drag Force

This is the force an object experiences while travelling through a medium. When something travels through the air it needs to displace air as it travels and because of this the air exerts a force on the object. This becomes an important force when you move fast and a lot of thought is taken to try and reduce the amount of drag force a sports car experiences.

The drag force is very useful for parachutists. They jump from high altitudes and if there was no drag force, then they would continue accelerating all the way to the ground. Parachutes are wide because the more surface area you show, the greater the drag force and hence the slower you hit the ground.

5.7 Summary of Important Quantities, Equations and Concepts

Equilibrium Objects at rest or moving with constant velocity are in *equilibrium* and have a *zero resultant force.*

Equilibrant The *equilibrant* of any number of forces is the single force required to produce equilibrium.

Triangle Law for Forces in Equilibrium Three forces in equilibrium can be represented in magnitude and direction by the three sides of a triangle taken in order.

Newton's First Law Every object will remain at rest or in uniform motion in a straight line unless it is made to change its state by the action of an *external force.*

Newton's Second Law The resultant force acting on a body results in an acceleration which is in the same direction as the resultant force and is directly proportional to the magnitude of this force and inversely proportional to the mass of the object.

Quantity	Symbol	S.I. Units	Direction
Mass	m	kg	—
Acceleration	$\overrightarrow{a}$	$m.s^{-2}$	✓
Force	$\overrightarrow{F}$	$kg.m.s^{-2}$ **or** N	✓

Table 5.1: Summary of the symbols and units of the quantities used in **Force**

Newton's Third Law For every force or action there is an *equal* but *opposite* force or reaction.

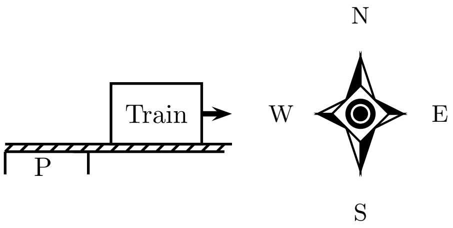
Chapter 6

Rectilinear Motion

6.1 What is rectilinear motion?

Rectilinear motion means motion along a straight line. This is a useful topic to study for learning how to describe the movement of cars along a straight road or of trains along straight railway tracks. In this section you have only 2 directions to worry about: (1) along the direction of motion, and (2) opposite to the direction of motion.

To illustrate this imagine a train heading east.



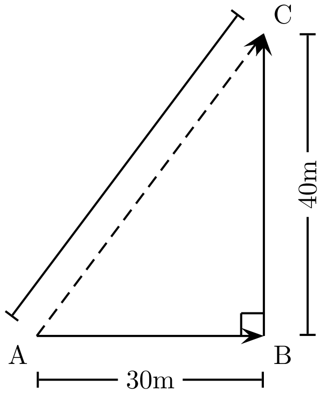
If it is accelerating away from the station platform (P), the direction of acceleration is the same as the direction of the train's velocity - east. If it is braking the direction of acceleration is opposite to the direction of its motion, i.e. west.

6.2 Speed and Velocity

Let's take a moment to review our definitions of velocity and speed by looking at the worked example below:

Worked Example 24

Speed and Velocity



Question: A cyclist moves from A through B to C in 10 seconds. Calculate both his speed and his velocity.

Answer:

Step 1 : Decide what information has been supplied

The question explicitly gives

- the distance between A and B
- the distance between B and C
- the total time for the cyclist to go from A through B to C

all in the correct units!

Remember to check the units!

Step 2 : Determine the cyclist's speed

His speed - a scalar - will be

$$\begin{aligned} v &= \frac{s}{t} \\ &= \frac{30m + 40m}{10s} \\ &= 7\frac{m}{s} \end{aligned}$$

Step 3 : First determine the cyclist's resultant displacement

Since velocity is a vector we will first need to find the resultant displacement of the cyclist. His velocity will be

$$\overrightarrow{v} = \frac{\overrightarrow{s}}{t}$$

The total displacement is the vector from A to C, and this is just the resultant of the two displacement vectors, ie.

$$\overrightarrow{s} = \overrightarrow{AC} = \overrightarrow{AB} + \overrightarrow{BC}$$

Using the rule of Pythagoras:

$$\begin{aligned}\overrightarrow{s} &= \sqrt{(30m)^2 + (40m)^2} \\ &= 50m\ in\ the\ direction\ from\ A\ to\ C\end{aligned}$$

Step 4 : Now we can determine the average velocity from the displacement and the time

$$\begin{aligned}\therefore \overrightarrow{v} &= \frac{50m}{10s} \\ &= 5\frac{m}{s}\ in\ the\ direction\ from\ A\ to\ C\end{aligned}$$

For this cyclist, his velocity is not the same as his speed because there has been a change in the direction of his motion. If the cyclist traveled directly from A to C *without* passing through B his speed would be

$$\begin{aligned}v &= \frac{50m}{10s} \\ &= 5\frac{m}{s}\end{aligned}$$

and his velocity would be

$$\begin{aligned}\overrightarrow{v} &= \frac{50m}{10s} \\ &= 5\frac{m}{s}\ in\ the\ direction\ from\ A\ to\ C\end{aligned}$$

In this case where the cyclist is not undergoing any change of direction (ie. he is traveling in a straight line) the magnitudes of the speed and the velocity are the same. This is the defining principle of rectilinear motion.

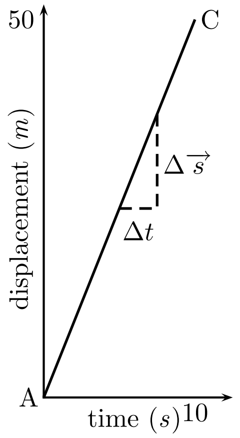
Important: For motion along a *straight line* the magnitudes of speed and velocity are the same, and the magnitudes of the distance and displacement are the same.

6.3 Graphs

In physics we often use graphs as important tools for picturing certain concepts. Below are some graphs that help us picture the concepts of displacement, velocity and acceleration.

6.3.1 Displacement-Time Graphs

Below is a graph showing the displacement of the cyclist from A to C:



This graphs shows us how, in 10 seconds time, the cyclist has moved from A to C. We know the gradient (slope) of a graph is defined as the change in y divided by the change in x, i.e $\frac{\Delta y}{\Delta x}$. In this graph the gradient of the graph is just $\frac{\Delta\overrightarrow{s}}{\Delta t}$ - and this is just the expression for velocity.

Important: The slope of a displacement-time graph gives the velocity.

The slope is the same all the way from A to C, so the cyclist's velocity is constant over the entire displacement he travels.

In figure 6.1 are examples of the displacement-time graphs you will encounter.

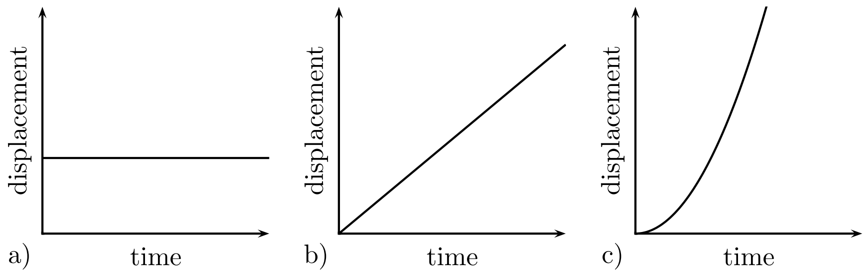


Figure 6.1: Some common displacement-time graphs:

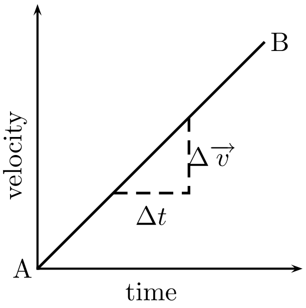
a) shows the graph for an object stationary over a period of time. The gradient is zero, so the object has zero velocity.

b) shows the graph for an object moving at a constant velocity. You can see that the displacement is increasing as time goes on. The gradient, however, stays constant (remember: its the slope of a straight line), so the velocity is constant. Here the gradient is positive, so the object is moving in the direction we have defined as positive.

c) shows the graph for an object moving at a constant acceleration. You can see that both the displacement and the velocity (gradient of the graph) increase with time. The gradient is increasing with time, thus the velocity is increasing with time and the object is accelerating.

6.3.2 Velocity-Time Graphs

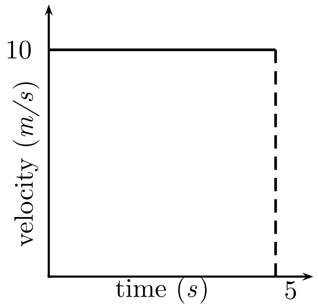
Look at the velocity-time graph below:



This is the velocity-time graph of a cyclist traveling from A to B at a constant acceleration, i.e. with steadily increasing velocity. The gradient of this graph is just $\frac{\Delta\overrightarrow{v}}{\Delta t}$ - and this is just the expression for acceleration. Because the slope is the same at all points on this graph, the acceleration of the cyclist is constant.

Important: The slope of a velocity-time graph gives the acceleration.

Not only can we get the acceleration of an object from its velocity-time graph, but we can also get some idea of the displacement traveled. Look at the graph below:



This graph shows an object moving at a constant velocity of $10m/s$ for a duration of $5s$. The area between the graph and the time axis (the (NOTE TO SELF: SHADED) area) of the above plot will give us the displacement of the object during this time. In this case we just need to calculate the area of a rectangle with width $5s$ and height $10m/s$

$$\begin{aligned}\text{area of rectangle} &= \text{height} \times \text{width} \\ &= \overrightarrow{v} \times t \\ &= 10\frac{m}{s} \times 5s \\ &= 50m \\ &= \overrightarrow{s} = \text{displacement}\end{aligned}$$

So, here we've shown that an object traveling at $10m/s$ for $5s$ has undergone a displacement of $50m$.

Important: The area between a velocity-time graph and the 'time' axis gives the displacement of the object.

Here are a couple more velocity-time graphs to get used to:

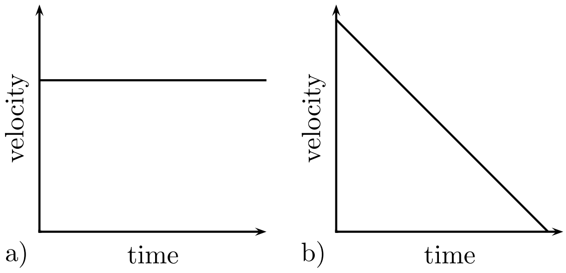


Figure 6.2: Some common velocity-time graphs:

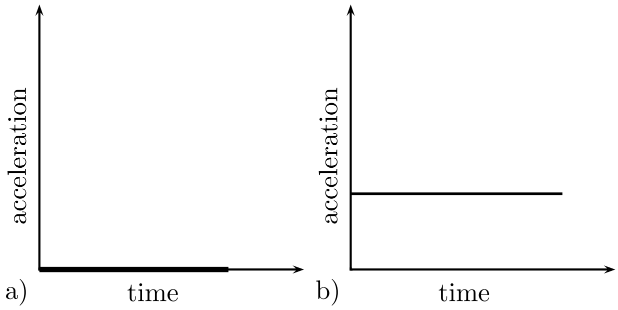
In figure 6.2 are examples of the displacement-time graphs you may encounter.

a) shows the graph for an object moving at a constant velocity over a period of time. The gradient is zero, so the object is not accelerating.

b) shows the graph for an object which is decelerating. You can see that the velocity is decreasing with time. The gradient, however, stays constant (remember: its the slope of a straight line), so the acceleration is constant. Here the gradient is negative, so the object is accelerating in the opposite direction to its motion, hence it is decelerating.

6.3.3 Acceleration-Time Graphs

In this chapter on rectilinear motion we will only deal with objects moving at a constant acceleration, thus all acceleration-time graphs will look like these two:

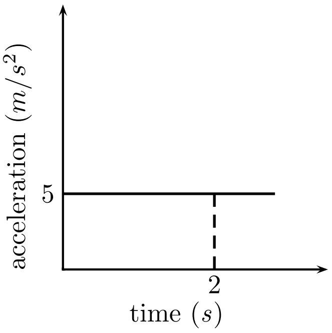


Here is a description of the graphs below:

a) shows the graph for an object which is either stationary or traveling at a constant velocity. Either way, the acceleration is zero over time.

b) shows the graph for an object moving at a constant acceleration. In this case the acceleration is positive - remember that it can also be negative.

We can obtain the velocity of a particle at some given time from an acceleration time graph - it is just given by the area between the graph and the time-axis. In the graph below, showing an object at a constant positive acceleration, the increase in velocity of the object after 2 seconds corresponds to the (NOTE TO SELF: shaded) portion.



$$\begin{aligned}\text{area of rectangle} &= \overrightarrow{a} \times t \\ &= 5\frac{m}{s^2} \times 2s \\ &= 10\frac{m}{s} \\ &= \overrightarrow{v}\end{aligned}$$

Its useful to remember the set of graphs below when working on problems. Figure 6.3 shows how displacement, velocity and time relate to each other. Given a displacement-time graph like the one on the left, we can plot the corresponding velocity-time graph by remembering that the slope of a displacement-time graph gives the velocity. Similarly, we can plot an acceleration-time graph from the gradient of the velocity-time graph.

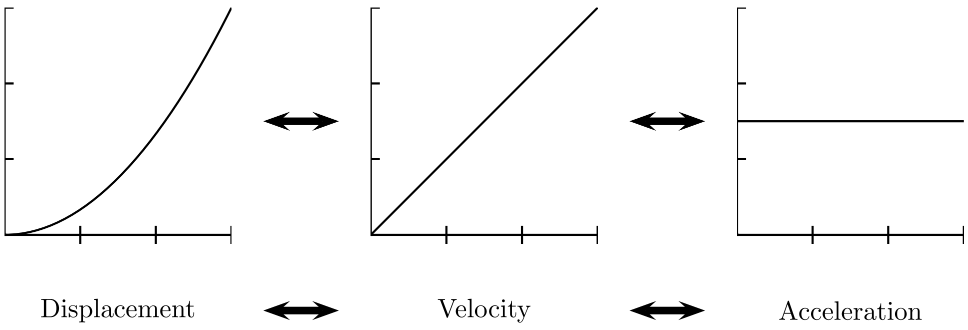


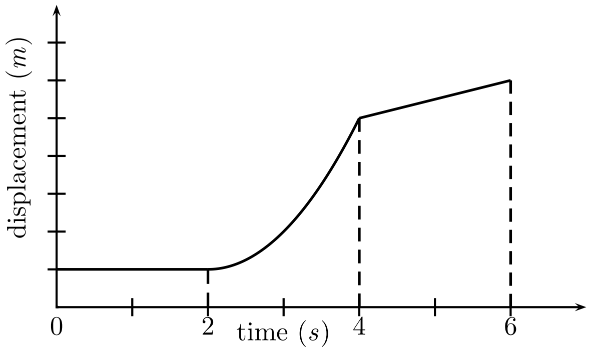
Figure 6.3: A Relationship Between Displacement, Velocity and Acceleration

6.3.4 Worked Examples

Worked Example 25

Relating displacement-, velocity-, and acceleration-time graphs

Question: Given the displacement-time graph below, draw the corresponding velocity-time and acceleration-time graphs, and then describe the motion of the object.



Answer:
Step 1 : Decide what information is supplied
The question explicitly gives a displacement-time graph.
Step 2 : Decide what is asked?
3 things are required:

1. Draw a velocity-time graph
2. Draw an acceleration-time graph

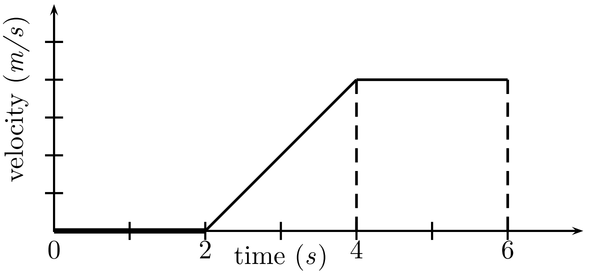
3. Describe the behaviour of the object

Step 3 : Velocity-time graph - 0-2 seconds
For the first 2 seconds we can see that the displacement remains constant - so the object is not moving, thus it has zero velocity during this time. We can reach this conclusion by another path too: remember that the gradient of a displacement-time graph is the velocity. For the first 2 seconds we can see that the displacement-time graph is a horizontal line, ie. it has a gradient of zero. Thus the velocity during this time is zero and the object is stationary.

Step 4 : Velocity-time graph - 2-4 seconds
For the next 2 seconds, displacement is increasing with time so the object is moving. Looking at the gradient of the displacement graph we can see that it is not constant. In fact, the slope is getting steeper (the gradient is increasing) as time goes on. Thus, remembering that the gradient of a displacement-time graph is the velocity, the velocity must be increasing with time during this phase.

Step 5 : Velocity-time graph - 4-6 seconds
For the final 2 seconds we see that displacement is still increasing with time, but this time the gradient is constant, so we know that the object is now travelling at a constant velocity, thus the velocity-time graph will be a horizontal line during this stage.

So our velocity-time graph looks like this one below. Because we haven't been given any values on the vertical axis of the displacement-time graph, we cannot figure out what the exact gradients are and hence what the values of the velocity are. In this type of question it is just important to show whether velocities are positive or negative, increasing, decreasing or constant.



Once we have the velocity-time graph its much easier to get the acceleration-time graph as we know that the gradient of a velocity-time graph is the just the acceleration.

Step 6 : Acceleration-time graph - 0-2 seconds
For the first 2 seconds the velocity-time graph is horizontal at zero, thus it has a gradient of zero and there is no acceleration during this time. (This makes sense because we know from the displacement time graph that the object is stationary during this time, so it can't be accelerating).

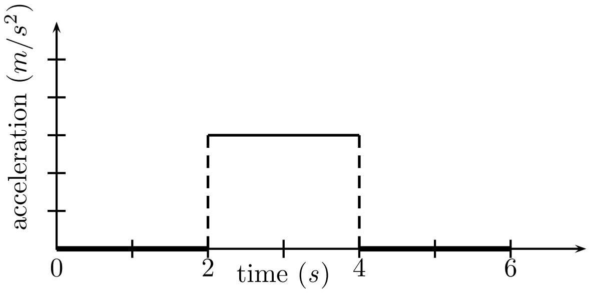
Step 7 : Acceleration-time graph - 2-4 seconds
For the next 2 seconds the velocity-time graph has a positive gradient. This gradient is not changing (i.e. its constant) throughout these 2 seconds so there must be a

constant positive acceleration.

Step 8 : Accleration-time graph - 4-6 seconds

For the final 2 seconds the object is traveling with a constant velocity. During this time the gradient of the velocity-time graph is once again zero, and thus the object is not accelerating.

The acceleration-time graph looks like this:



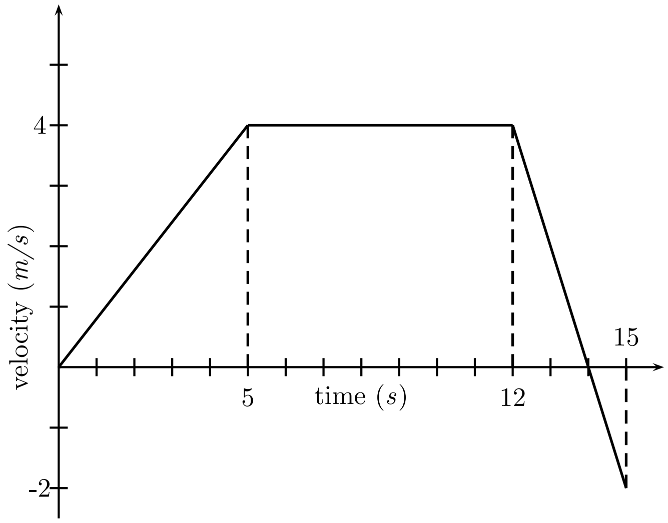
Step 9 : A description of the object's motion

A brief description of the motion of the object could read something like this: At $t = 0s$ and object is stationary at some position and remains stationary until $t = 2s$ when it begins accelerating. It accelerates in a positive direction for 2 seconds until $t = 4s$ and then travels at a constant velocity for a further 2 seconds.

Worked Example 26

Calculating distance from a velocity-time graph

Question: The velocity-time graph of a car is plotted below. Calculate the displacement of the car has after 15 seconds.



Answer:

Step 1 : Decide how to tackle the problem

We are asked to calculate the displacement of the car. All we need to remember here is that the area between the velocity-time graph and the time axis gives us the displacement.

Step 2 : Determine the area under the velocity-time graph

For $t = 0s$ to $t = 5s$ this is the triangle on the left:

$$\begin{aligned} Area\triangle &= \frac{1}{2}b \times h \\ &= \frac{1}{2}5s \times 4m/s \\ &= 10m \end{aligned}$$

For $t = 5s$ to $t = 12s$ the displacement is equal to the area of the rectangle

$$\begin{aligned} Area\square &= w \times h \\ &= 7s \times 4m/s \\ &= 28m \end{aligned}$$

For $t = 12s$ to $t = 14s$ the displacement is equal to the area of the triangle above the time axis on the right

$$\begin{aligned} Area\triangle &= \frac{1}{2}b \times h \\ &= \frac{1}{2}2s \times 4m/s \\ &= 4m \end{aligned}$$

For $t = 14s$ to $t = 15s$ the displacement is equal to the area of the triangle below the time axis

$$\begin{aligned} Area\triangle &= \frac{1}{2}b \times h \\ &= \frac{1}{2}1s \times 2m/s \\ &= 1m \end{aligned}$$

Step 3 : Determine the total displacement of the car
Now the total displacement of the car is just the sum of all of these areas. HOWEVER, because in the last second (from $t = 14s$ to $t = 15s$) the velocity of the car is negative, it means that the car was going in the opposite direction, i.e. back where it came from! So, to get the total displacement, we have to add the first 3 areas (those with positive displacements) and subtract the last one (because it signifies a displacement in the opposite direction).

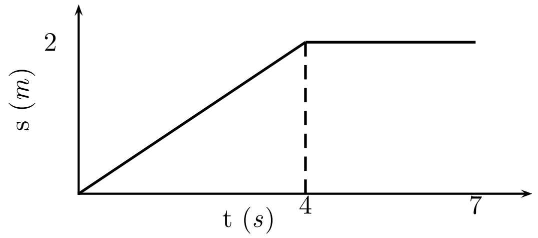
$$\begin{aligned} \overrightarrow{s} &= 10 + 28 + 4 - 1 \\ &= 41m\ in\ the\ positive\ direction \end{aligned}$$

Worked Example 27

Velocity from a displacement-time graph

Question: Given the diplacement-time graph below,

1. what is the velocity of the object during the first 4 seconds?
2. what is the velocity of the object from $t = 4s$ to $t = 7s$?



Answer:
Step 1 : The velocity during the first 4 seconds
The velocity is given by the slope of a displacement-time graph. During the first 4 seconds, this is

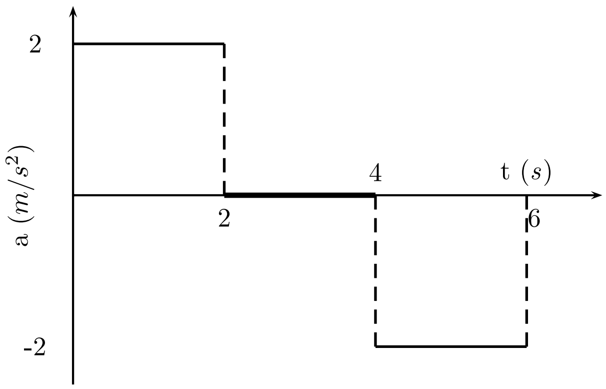
$$\begin{aligned} \overrightarrow{v} &= \frac{\Delta s}{\Delta t} \\ &= \frac{2m}{4s} \\ &= 0.5m/s \end{aligned}$$

Step 2 : The velocity during the last 3 seconds
For the last 3 seconds we can see that the displacement stays constant, and that the gradient is zero. Thus $\overrightarrow{v} = 0 m/s$

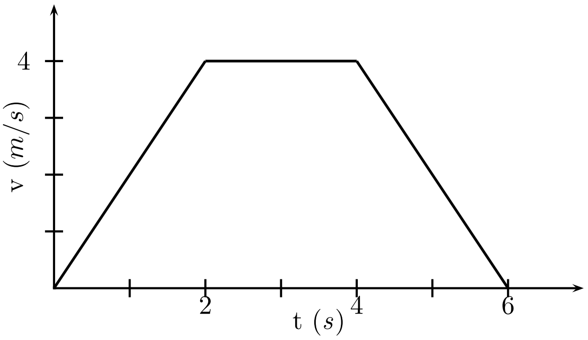
Worked Example 28

From an acceleration-time graph to a velocity-time graph

Question: Given the acceleration-time graph below, assume that the object starts from rest and draw its velocity-time graph.



Answer: Once again attempt to draw the graph in time sections, i.e. first draw the velocity-time graph for the first 2 seconds, then for the next 2 seconds and so on.



6.4 Equations of Motion

This section is about solving problems relating to uniformly accelerated motion. We'll first introduce the variables and the equations, then we'll show you how to derive them, and after that we'll do a couple of examples.

$$\begin{aligned}
u &= \text{starting velocity (m/s) at } t = 0 \\
v &= \text{final velocity (m/s) at time } t \\
s &= \text{displacement (m)} \\
t &= \text{time (s)} \\
a &= \text{acceleration (m/s}^2\text{)}
\end{aligned}$$

$$\begin{aligned}
v &= u + at && (6.1) \\
s &= \frac{(u+v)}{2}t && (6.2) \\
s &= ut + \frac{1}{2}at^2 && (6.3) \\
v^2 &= u^2 + 2as && (6.4)
\end{aligned}$$

Make sure you can rhyme these off, they are *very* important! There are so many different types of questions for these equations. Basically when you are answering a question like this:

1. Find out what values you have and write them down.
2. Figure out which equation you need.
3. *Write it down!!!*
4. Fill in all the values you have and get the answer.

Interesting Fact: Galileo Galilei of Pisa, Italy, was the first to determined the correct mathematical law for acceleration: the total distance covered, starting from rest, is proportional to the square of the time. He also concluded that objects retain their velocity unless a force – often friction – acts upon them, refuting the accepted Aristotelian hypothesis that objects "naturally" slow down and stop unless a force acts upon them. This principle was incorporated into Newton's laws of motion (1st law).

Equation 6.1

By the definition of acceleration

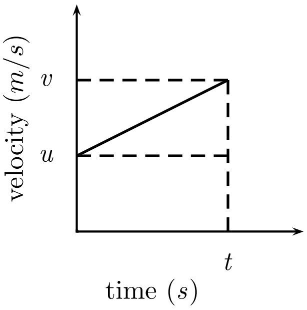
$$a = \frac{\Delta v}{t}$$

where Δv is the change in velocity, i.e. $\Delta v = v - u$. Thus we have

$$\begin{aligned} a &= \frac{v-u}{t} \\ v &= u + at \end{aligned}$$

Equation 6.2

In the previous section we saw that displacement can be calculated from the area between a velocity-time graph and the time-axis. For *uniformly accelerated motion* the most complicated velocity-time graph we can have is a straight line. Look at the graph below - it represents an object with a starting velocity of u, accelerating to a final velocity v over a total time t.



To calculate the final displacement we must calculate the area under the graph - this is just the area of the rectangle added to the area of the triangle. (NOTE TO SELF: SHADING)

$$\begin{aligned} Area\triangle &= \frac{1}{2}b \times h \\ &= \frac{1}{2}t \times (v-u) \\ &= \frac{1}{2}vt - \frac{1}{2}ut \end{aligned}$$

$$\begin{aligned} Area\square &= w \times h \\ &= t \times u \\ &= ut \end{aligned}$$

$$\begin{aligned} Displacement &= Area\square + Area\triangle \\ s &= ut + \frac{1}{2}vt - \frac{1}{2}ut \\ &= \frac{(u+v)}{2}t \end{aligned}$$

Equation 6.3

This equation is simply derived by eliminating the final velocity v in equation 6.2. Remembering from equation 6.1 that

$$v = u + at$$

then equation 6.2 becomes

$$\begin{aligned} s &= \frac{u+u+at}{2}t \\ &= \frac{2ut+at^2}{2} \\ &= ut+\frac{1}{2}at^2 \end{aligned}$$

Equation 6.4

This equation is just derived by eliminating the time variable in the above equation. From Equation 6.1 we know

$$t = \frac{v-u}{a}$$

Substituting this into Equation 6.3 gives

$$\begin{aligned} s &= u(\frac{v-u}{a}) + \frac{1}{2}a(\frac{v-u}{a})^2 \\ &= \frac{uv}{a} - \frac{u^2}{a} + \frac{1}{2}a(\frac{v^2-2uv+u^2}{a^2}) \\ &= \frac{uv}{a} - \frac{u^2}{a} + \frac{v^2}{2a} - \frac{uv}{a} + \frac{u^2}{2a} \\ 2as &= -2u^2+v^2+u^2 \\ v^2 &= u^2+2as \qquad (6.5) \end{aligned}$$

This gives us the final velocity in terms of the initial velocity, acceleration and displacement and is independent of the time variable.

Worked Example 29

Question: A racing car has an initial velocity of $100m/s$ and it covers a displacement of $725m$ in $10s$. Find its acceleration.

Answer:

Step 1 : Decide what information has been supplied

We are given the quantities u, s and t - all in the correct units. We need to find a.

Step 2 : Find an equation of motion relating the given information to the acceleration

We can use equation 6.3

$$s = ut + \frac{1}{2}at^2$$

Step 3 : Rearrange the equation if needed

We want to determine the acceleration so we rearrane equation 6.3 to put acceleration on the left of the equals sign:

$$a = \frac{2(s-ut)}{t^2}$$

Step 4 : Do the calculation
Substituting in the values of the known quantities this becomes

$$\begin{aligned} a &= \frac{2(725m - 100\frac{m}{s}\cdot 10s)}{10^2 s^2} \\ &= \frac{2(-275m)}{100s^2} \\ &= -5.5\frac{m}{s^2} \end{aligned}$$

Step 5 : Quote the final answer
The racing car is accelerating at -5.5$\frac{m}{s^2}$, or we could say it is decelerating at 5.5$\frac{m}{s^2}$.

Worked Example 30

Question: An object starts from rest, moves in a straight line with a constant acceleration and covers a distance of $64m$ in $4s$. Calculate

- its acceleration
- its final velocity
- at what time the object had covered half the total distance
- what distance the object had covered in half the total time.

Answer:
Step 1 : Decide what information is supplied
We are given the quantities u, s and t in the correct units.
Step 2 : Acceleration: Find an equation to calculate the acceleration and rearrange
To calculate the acceleration we can use equation 6.3.
Step 3 : Rearrange to make a subject of the formula

$$a = \frac{2(s - ut)}{t^2}$$

Step 4 : Do the calculation
Substituting in the values of the known quantities this becomes

$$\begin{aligned} a &= \frac{2(64m - 0\frac{m}{s}4s}{4^2 s^2} \\ &= \frac{128m}{16s^2} \\ &= 8\frac{m}{s^2} \end{aligned}$$

Step 5 : Final velocity: Find an equation to calculate the final velocity
We can use equation 6.1 - remember we now also know the acceleration of the object.

$$v = u + at$$

Step 6 : Do the calculation

$$\begin{aligned} v &= 0\frac{m}{s} + (8\frac{m}{s^2})(4s) \\ &= 32\frac{m}{s} \end{aligned}$$

Step 7 : Time at half the distance: Find an equation to relate the unknown and known quantities
Here we have the quantities s, u and a so we do this in 2 parts, first using equation 6.4 to calculate the velocity at half the distance, i.e. $32m$:

$$\begin{aligned} v^2 &= u^2 + 2as \\ &= (0m)^2 + 2(8m/s^2)(32m) \\ &= 512m^2/s^2 \\ v &= 22.6m/s \end{aligned}$$

Now we can use equation 6.2 to calculate the time at this distance:

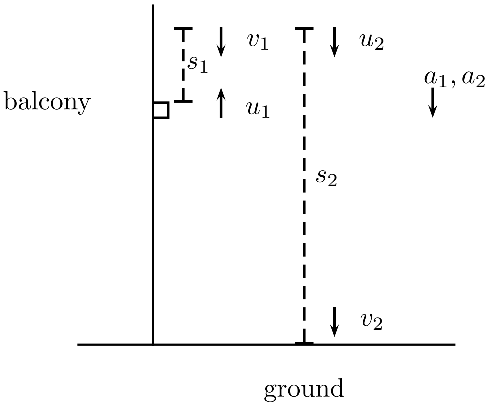
$$\begin{aligned} t &= \frac{2s}{u+v} \\ &= \frac{(2)(32m)}{0m/s + 22.6m/s} \\ &= 2.8s \end{aligned}$$

Step 8 : Distance at half the time: Find an equation to relate the distance and time
To calculate the distance the object has covered in half the time. Half the time is $2s$. Thus we have u, a and t - all in the correct units. We can use equation 6.3 to get the distance:

$$\begin{aligned} s &= ut + \frac{1}{2}at^2 \\ &= (0m/s)(2s) + \frac{1}{2}(8\frac{m}{s^2})(2s)^2 \\ &= 16m \end{aligned}$$

Worked Example 31

Question: A ball is thrown vertically upwards with a velocity of $10m/s$ from the balcony of a tall building. The balcony is $15m$ above the ground and gravitational accleration is $10m/s^2$. Find a) the time required for the ball to hit the ground, and b) the velocity with which it hits the ground.
Answer:
Step 1 : Draw a rough sketch of the problem
In most cases helps to make the problem easier to understand if we draw ourselves a picture like the one below:



where the subscript 1 refers to the upward part of the ball's motion and the subscript 2 refers to the downward part of the ball's motion.

Step 2 : Decide how to tackle the problem

First the ball goes upwards with gravitational acceleration slowing it until it reaches its highest point - here its speed is $0m/s$ - then it begins descending with gravitational acceleration causing it to increase its speed on the way down. We can separate the motion into 2 stages:

Stage 1 - the upward motion of the ball

Stage 2 - the downward motion of the ball.

We'll choose the upward direction as positive - this means that gravitation acceleraton is negative. We have used this and we'll begin by solving for all the variables of Stage 1.

Step 3 : For Stage 1, decide what information is given

We have these quantities:

$$\begin{aligned} u_1 &= 10m/s \\ v_1 &= 0m/s \\ a_1 &= -10m/s^2 \\ t_1 &= ? \\ s_1 &= ? \end{aligned}$$

Step 4 : Find the time for stage 1

Using equation 6.1 to find t_1:

$$\begin{aligned} v_1 &= u_1 + a_1 t_1 \\ t_1 &= \frac{v_1 - u_1}{a_1} \\ &= \frac{0m/s - 10m/s}{-10m/s^2} \\ &= 1s \end{aligned}$$

Step 5 : Find the distance travelled during stage 1

We can find s_1 by using equation 6.4

$$\begin{aligned} v_1^2 &= u_1^2 + 2a_1 s_1 \\ s_1 &= \frac{v_1^2 - u_1^2}{2a} \\ &= \frac{(0m/s)^2 - (10m/s)^2}{2(-10m/s^2)} \\ &= 5m \end{aligned}$$

Step 6 : For Stage 2, decide what information is supplied
For Stage 2 we have the following quantities:

$$\begin{aligned} u_2 &= 0m/s \\ v_2 &= ? \\ a_2 &= -10m/s^2 \\ t_2 &= ? \\ s_2 &= -15m - 5m = 20m \end{aligned}$$

Step 7 : Determine the velocity at the end of stage 2
We can determine the final velocity v_2 using equation 6.4:

$$\begin{aligned} v_2^2 &= u_2^2 + 2a_2 s_2 \\ &= (0m/s)^2 + 2(-10m/s^2)(-20m) \\ &= 400(m/s)^2 \\ v_2 &= 20m/s \text{ downwards} \end{aligned}$$

Step 8 : Determind the time for stage 2
Now we can determine the time for Stage 2, t_2, from equation 6.1:

$$\begin{aligned} v_2 &= u_2 + a_2 t_2 \\ t_2 &= \frac{v_2 - u_2}{a_2} \\ &= \frac{-20m/s - 0m/s}{-10m/s^2} \\ &= 2s \end{aligned}$$

Step 9 : Quote the answers to the problem
Finally,
a) the time required for the stone to hit the ground is $t = t_1 + t_2 = 1s + 2s = 3s$
b) the velocity with which it hits the ground is just $v_2 = -20m/s$

These questions do not have the working out in them, but they are all done in the manner described on the previous page.

Question: A car starts off at 10 m/s and accelerates at 1 m/s^2 for 10 seconds. What is it's final velocity?
Answer: 20 m/s

Question: A car starts from rest, and accelerates at 1 m/s^2 for 10 seconds. How far does it move?
Answer: 50 m

Question: A car is going 30 m/s and stops in 2 seconds. What is it's stopping distance for this speed?
Answer: 30 m

Question: A car going at 20 m/s stops in a distance of 20 m/s.

1. What is it's deceleration?
2. If the car is 1 Tonne (1000 Kg, or 1 Mg) how much force do the brakes exert?

6.5 Important Equations and Quantities

Units			
Quantity	Symbol	Unit	Base S.I. Units
Displacement	$\overrightarrow{s}$	-	m + direction
Velocity	$\overrightarrow{u},\overrightarrow{v}$	-	$m.s^{-1}$ + direction
Distance	s	-	m
Speed	v	-	$m.s^{-1}$
Acceleration	$\overrightarrow{a}$	-	$m.s^{-1}$ + direction

Table 6.1: Units used in **Rectilinear Motion**

Chapter 7

Momentum

7.1 What is Momentum?

Momentum is a physical quantity which is closely related to forces. We will learn about this connection a little later. Remarkably momentum is a conserved quantity. This makes momentum extremely useful in solving a great variety of real-world problems. Firstly we must consider the definition of momentum.

Definition: The *momentum* of an object is defined as its mass multiplied by its velocity.

Mathematically,

$$\overrightarrow{p} = m\overrightarrow{v}$$

$\overrightarrow{p}$	: momentum ($kg.m.s^{-1}$ + direction)
m	: mass (kg)
$\overrightarrow{v}$	: velocity ($m.s^{-1}$ + direction)

Thus, momentum is a property of a moving object and is determined by its velocity and mass. A large truck travelling slowly can have the same momentum as a much smaller car travelling relatively fast.

Note the arrows in the equation defining momentum– momentum is a vector with the same direction as the velocity of the object.

Momentum is a vector with the same direction as the velocity.

Since the direction of an object's momentum is given by the direction of its motion, one can calculate an object's momentum in two steps:

- calculate the magnitude of the object's momentum using,

$$p = mv$$

p	: magnitude of momentum ($kg.m.s^{-1}$)
m	: mass (kg)
v	: magnitude of velocity ($m.s^{-1}$)

- include in the final answer the direction of the object's motion

Worked Example 32

Calculating Momentum 1

Question: A ball of mass $3kg$ moves at $2m.s^{-1}$ to the right. Calculate the ball's momentum.
Answer:
Step 1 : Decide what information has been supplied
The question explicitly gives

- the ball's mass, and
- the ball's velocity

in the correct units!

Remember to check the units!

Step 2 : Decide how to tackle the problem
What is being asked? We are asked to calculate the ball's momentum. From the definition of momentum,

$$\overrightarrow{p} = m\overrightarrow{v},$$

we see that we need the mass and velocity of the ball, which we are given.
Step 3 : Do the calculation
We calculate the magnitude of the ball's momentum,

$$\begin{aligned} p &= mv \\ &= (3kg)(2m.s^{-1}) = 6\ kg.m.s^{-1}. \end{aligned}$$

Step 4 : Quote the final answer
We quote the answer with the direction of the ball's motion included,

$$\overrightarrow{p} = 6\ kg.m.s^{-1}\ \textbf{to the right}$$

Worked Example 33

Calculating Momentum 2

Question: A ball of mass $500g$ is thrown at $2m.s^{-1}$. Calculate the ball's momentum.
Answer:
Step 1 : Decide what information is supplied
The question explicitly gives

- the ball's mass, and
- the magnitude of the ball's velocity

but with the ball's mass in the incorrect units!

Remember to check the units!

Step 2 : Decide how to tackle the problem
What is being asked? We are asked to calculate the momentum which is defined as

$$\overrightarrow{p} = m\overrightarrow{v}.$$

Thus, we need the mass and velocity of the ball but we have only its mass and the magnitude of its velocity. In order to determine the velocity of the ball we need the direction of the ball's motion. If the problem does not give an explicit direction we are forced to be general. In a case like this we could say that the direction of the velocity is *in the direction of motion of the ball.* This might sound silly but the lack of information in the question has forced us to be, and we are certainly not wrong! The ball's velocity is then $2m.s^{-1}$ in the direction of motion.
Step 3 : Convert the mass to the correct units

$$\begin{aligned} 1000g &= 1kg \\ 1 &= \frac{1kg}{1000g} \\ 500g \times 1 &= 500g \times \frac{1kg}{1000g} \\ &= 0.500kg \end{aligned}$$

Step 4 : Do the calculation
Now, let us find the magnitude of the ball's momentum,

$$\begin{aligned} p &= mv \\ &= (0.500kg)(2m.s^{-1}) = 1\ kg.m.s^{-1} \end{aligned}$$

Step 5 : Quote the final answer
Remember to include the direction of the momentum:

$$\overrightarrow{p} = 1\ kg.m.s^{-1}\ \textbf{in the direction of motion of the ball}$$

Worked Example 34

Calculating the Momentum of the Moon

Question: The moon is 384 400km away from the earth and orbits the earth in 27.3 days. If the moon has a mass of $7.35 \times 10^{22} kg$[1] what is the magnitude of its momentum if we assume a circular orbit?
Answer:
Step 1 : Decide what information has been supplied
The question explicitly gives

[1] This is $\frac{1}{81}$ of the mass of the earth

- the moon's mass,
- the distance to the moon, and
- the time for one orbit of the moon

with mass in the correct units but all other quantities in the incorrect units. The units we require are

Remember to check the units!

- seconds (s) for time, and
- metres (m) for distance

Step 2 : Decide how to tackle the problem
What is being asked? We are asked to calculate only the magnitude of the moon's momentum (i.e. we do not need to specify a direction). In order to do this we require the moon's mass and the magnitude of its velocity, since

$$p = mv.$$

Step 3 : Find the speed or magnitude of the moon's velocity
Speed is defined as,

$$speed \quad = \quad \frac{Distance}{time}$$

We are given the time the moon takes for one orbit but not how far it travels in that time. However, we can work this out from the distance to the moon and the fact that the moon's orbit is circular. Firstly let us convert the distance to the moon to the correct units,

$$\begin{aligned}
1km &= 1000m \\
1 &= \frac{1000m}{1km} \\
384\,400km \times 1 &= 384\,400km \times \frac{1000m}{1km} \\
&= 384\,400\,000m \\
&= 3.844 \times 10^8\ m
\end{aligned}$$

Using the equation for the circumference, C, of a circle in terms of its radius, we can determine the distance travelled by the moon in one orbit:

$$\begin{aligned}
C &= 2\pi r \\
&= 2\pi(3.844 \times 10^8\ m) \\
&= 2.42 \times 10^9\ m.
\end{aligned}$$

Next we must convert the orbit time, T, into the correct units. Using the fact that a day contains 24 hours, an hour consists of 60 minutes, and a minute is 60 seconds long,

$$\begin{aligned}
1day &= (24)(60)(60)seconds \\
1 &= \frac{(24)(60)(60)s}{1day} \\
27.3days \times 1 &= 27.3days\frac{(24)(60)(60)s}{1day} \\
&= 2.36 \times 10^6 s
\end{aligned}$$

Therefore,

$$T = 2.36 \times 10^6 s.$$

Combining the distance travelled by the moon in an orbit and the time taken by the moon to complete one orbit, we can determine the magnitude of the moon's velocity or speed,

$$\begin{aligned} v &= \frac{Distance}{time} \\ &= \frac{C}{T} \\ &= 1.02 \times 10^3 \; m.s^{-1}. \end{aligned}$$

Step 4 : Finally calculate the momentum and quote the answer
The magnitude of the moon's momentum is:

$$\begin{aligned} p &= mv \\ &= (7.35 \times 10^{22} kg)(1.02 \times 10^3 \; m.s^{-1}) \\ &= 7.50 \times 10^{25} \; kg.m.s^{-1}. \end{aligned}$$

7.2 The Momentum of a System

In Chapter **??** the concept of a system was introduced. The bodies that make up a system can have different masses and can be moving with different velocities. In other words they can have different momenta.

Definition: The *total momentum of a system* is the sum of the momenta of each of the objects in the system.

Since momentum is a vector, the techniques of vector addition discussed in Chapter **??** must be used to calculate the total momentum of a system. Let us consider an example.

Worked Example 35

Calculating the Total Momentum of a System

Question: Two billiard balls roll towards each other. They each have a mass of $0.3kg$. Ball 1 is moving at $v_1 = 1 \; m.s^{-1}$ to the right, while ball 2 is moving at $v_2 = 0.8 \; m.s^{-1}$ to the left. Calculate the total momentum of the system.

Answer:
Step 1 : Decide what information is supplied
The question explicitly gives

- the mass of each ball,
- the velocity of ball 1, $\overrightarrow{v_1}$, and

- the velocity of ball 2, $\overrightarrow{v_2}$,

all in the correct units!

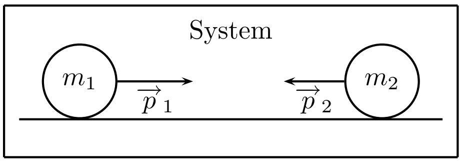
Remember to check the units!

Step 2 : Decide how to tackle the problem

What is being asked? We are asked to calculate the **total momentum of the system**. In this example our system consists of two balls. To find the total momentum we must sum the momenta of the balls,

$$\overrightarrow{p}_{total} = \overrightarrow{p_1} + \overrightarrow{p_2}$$

Since ball 1 is moving to the right, its momentum is in this direction, while the second ball's momentum is directed towards the left.



Thus, we are required to find the sum of two vectors acting along the same straight line. The algebraic method of vector addition introduced in Chapter **??** can thus be used.

Step 3 : Choose a positive direction

Let us choose right as the positive direction, then obviously left is negative.

Step 4 : Calculate the momentum

The total momentum of the system is then the sum of the two momenta taking the directions of the velocities into account. Ball 1 is travelling at 1 $m.s^{-1}$ *to the right* or +1 $m.s^{-1}$. Ball 2 is travelling at 0.8 $m.s^{-1}$ *to the left* or −0.8 $m.s^{-1}$. Thus,

Right is the positive direction

$$\begin{aligned}\overrightarrow{p}_{total} &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2}\\ &= (0.3kg)(+1\ m.s^{-1}) + (0.3kg)(-0.8\ m.s^{-1})\\ &= (+0.3\ kg.m.s^{-1}) + (-0.24\ kg.m.s^{-1})\\ &= +0.06\ kg.m.s^{-1}\\ &= 0.06\ kg.m.s^{-1}\ \textbf{to the right}\end{aligned}$$

In the last step the direction was added in words. Since the result in the second last line is positive, the total momentum of the system is in the positive direction (i.e. to the right).

7.3 Change in Momentum

If either an object's mass or velocity changes then its momentum too will change. If an object has an initial velocity $\overrightarrow{u}$ and a final velocity $\overrightarrow{v}$, then its change in momentum, $\Delta\overrightarrow{p}$, is

$$\boxed{\Delta\overrightarrow{p} \quad = \overrightarrow{p}_{final} - \overrightarrow{p}_{initial} = m\overrightarrow{v} - m\overrightarrow{u}}$$

Worked Example 36

Change in Momemtum

Question: A rubber ball of mass $0.8kg$ is dropped and strikes the floor at a velocity of 6 $m.s^{-1}$. It bounces back with an initial velocity of 4 $m.s^{-1}$. Calculate the change in momentum of the rubber ball caused by the floor.

Answer:

Step 1 : Decide what information has been supplied

The question explicitly gives

- the ball's mass,
- the ball's initial velocity, and
- the ball's final velocity

all in the correct units.

Do not be confused by the question referring to the ball bouncing back with an "initial velocity of 4 $m.s^{-1}$". The word "initial" is included here since the ball will obviously slow down with time and 4 $m.s^{-1}$ is the speed immediately after bouncing from the floor.

Remember to check the units!

Step 2 : Decide how to tackle the problem

What is being asked? We are asked to calculate the change in momentum of the ball,

$$\Delta\overrightarrow{p} \quad = \quad m\overrightarrow{v} - m\overrightarrow{u}.$$

We have everything we need to find $\Delta\overrightarrow{p}$. Since the initial momentum is directed downwards and the final momentum is in the upward direction, we can use the algebraic method of subtraction discussed in the vectors chapter.

Step 3 : Choose a positive direction

Let us choose down as the positive direction. Then substituting,

Down is the positive direction

Step 4 : Do the calculation and quote the answer

$$\begin{aligned}
\Delta\overrightarrow{p} &= m\overrightarrow{v} - m\overrightarrow{u} \\
&= (0.8kg)(-4\ m.s^{-1}) - (0.8kg)(+6\ m.s^{-1}) \\
&= (0.8kg)(-10\ m.s^{-1}) \\
&= -8\ kg.m.s^{-1} \\
&= 8\ kg.m.s^{-1}\ \textbf{up}
\end{aligned}$$

where we remembered in the last step to include the direction of the change in momentum in words.

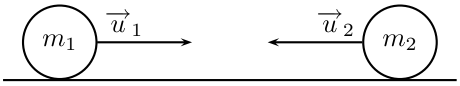


Figure 7.1: Before the collision.

7.4 What properties does momentum have?

You may at this stage be wondering why there is a need for introducing momentum. Remarkably momentum is a conserved quantity. Within an isolated system the total momentum is constant. No matter what happens to the individual bodies within an isolated system, the total momentum of the system never changes! Since momentum is a vector, its conservation implies that both its magnitude and its direction remains the same.

This **Principle of Conservation of Linear Momentum** is one of the most fundamental principles of physics and it alone justifies the definition of momentum. Since momentum is related to the motion of objects, we can use its conservation to make predictions about what happens in collisions and explosions. If we bang two objects together, by conservation of momentum, the total momentum of the objects before the collision is equal to their total momentum after the collision.

Momentum is **conserved** in isolated system

Principle of Conservation of Linear Momentum:
The total linear momentum of an isolated system is constant.
or
In an isolated system the total momentum before a collision (or explosion) is equal to the total momentum after the collision (or explosion).

Let us consider a simple collision of two pool or billiard balls. Consider the first ball (mass m_1) to have an initial velocity ($\overrightarrow{u_1}$). The second ball (mass m_2) moves towards the first ball with an initial velocity $\overrightarrow{u_2}$. This situation is shown in Figure 7.1. If we add the momenta of each ball we get a total momentum for the system. This total momentum is then

$$\overrightarrow{p}_{total\ before} = m_1\overrightarrow{u_1} + m_2\overrightarrow{u_2},$$

After the two balls collide and move away they each have a different momentum. If we call the final velocity of ball 1 $\overrightarrow{v_1}$ and the final velocity of ball 2 $\overrightarrow{v_2}$ (see Figure 7.2), then the total momentum of the system after the collision is

$$\overrightarrow{p}_{total\ after} = m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2},$$

This system of two balls is isolated since there are no external forces acting on the balls. Therefore, by the principle of conservation of linear momentum, the total momentum before the collision is equal to the total momentum after the collision. This gives the equation for the conservation of momentum in a collision of two objects,

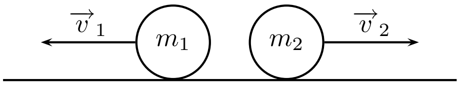


Figure 7.2: After the collision.

$$\overrightarrow{p}_{total\ before} = \overrightarrow{p}_{total\ after}$$

$$m_1\overrightarrow{u_1} + m_2\overrightarrow{u_2} = m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2}$$

m_1 : mass of object 1 (kg)
m_2 : mass of object 2 (kg)

$\overrightarrow{u_1}$: initial velocity of object 1 ($m.s^{-1}$ + direction)
$\overrightarrow{u_2}$: initial velocity of object 2 ($m.s^{-1}$ + direction)

$\overrightarrow{v_1}$: final velocity of object 1 ($m.s^{-1}$ + direction)
$\overrightarrow{v_2}$: final velocity of object 2 ($m.s^{-1}$ + direction)

This equation is always true- momentum is always conserved in collisions.

The chapter ‘Collisions and Explosions’ (Chapter **??**) deals with applications of momentum conservation.

Momentum is always conserv in collisions!

7.5 Impulse

At the beginning of this chapter it was mentioned that momentum is closely related to force. We will now explain the nature of this connection.

Consider an object of mass m moving with constant acceleration $\overrightarrow{a}$. During a time Δt the object's velocity changes from an initial velocity $\overrightarrow{u}$ to a final velocity $\overrightarrow{v}$ (refer to Figure 7.3). We know from Newton's First Law that there must be a resultant force $\overrightarrow{F}_{Res}$ acting on the object.

Starting from Newton's Second Law,

$$\begin{aligned}
\overrightarrow{F}_{Res} &= m\overrightarrow{a} \\
&= m(\frac{\overrightarrow{v} - \overrightarrow{u}}{\Delta t}) \qquad \text{since } \overrightarrow{a} = \frac{\overrightarrow{v} - \overrightarrow{u}}{\Delta \text{t}} \\
&= \frac{m\overrightarrow{v} - m\overrightarrow{u}}{\Delta t} \\
&= \frac{\overrightarrow{p}_{final} - \overrightarrow{p}_{initial}}{\Delta t} \\
&= \frac{\Delta\overrightarrow{p}}{\Delta t}
\end{aligned}$$

This alternative form of Newton's Second Law is called the **Law of Momentum**.

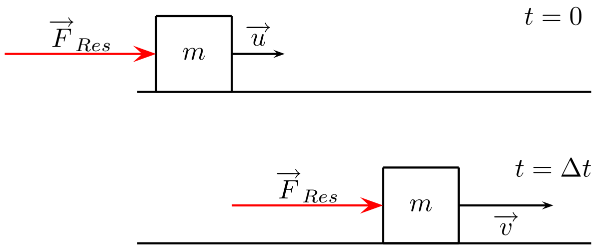


Figure 7.3: An object under the action of a resultant force.

Law of Momentum: The applied resultant force acting on an object is equal to the rate of change of the object's momentum and this force is in the direction of the change in momentum.

Mathematically,

$$\overrightarrow{F}_{Res} = \frac{\Delta\overrightarrow{p}}{\Delta t}$$

$\overrightarrow{F}_{Res}$: resultant force (N + direction)

$\Delta\overrightarrow{p}$: change in momentum ($kg.m.s^{-1}$ + direction)

Δt : time over which $\overrightarrow{F}_{Res}$ acts (s)

Rearranging the Law of Momentum,

$$\overrightarrow{F}_{Res}\Delta t \quad = \quad \Delta\overrightarrow{p}.$$

The product $\overrightarrow{F}_{Res}\Delta t$ is called impulse,

$$\boxed{\text{Impulse} \equiv \overrightarrow{F}_{Res}\Delta t = \Delta\overrightarrow{p}}$$

From this equation we see, that for a given change in momentum, $\overrightarrow{F}_{Res}\Delta t$ is fixed. Thus, if F_{Res} is reduced, Δt must be increased (i.e. the resultant force must be applied for longer). Alternatively if Δt is reduced (i.e. the resultant force is applied for a shorter period) then the resultant force must be increased to bring about the same change in momentum.

Worked Example 37

Impulse and Change in momentum

Question: A 150 N resultant force acts on a 300 kg object. Calculate how long it takes this force to change the object's velocity from 2 $m.s^{-1}$ *to the right* to 6 $m.s^{-1}$ *to the right*.

Answer:

Step 1 : Decide what information is supplied

The question explicitly gives

- the object's mass,
- the object's initial velocity,
- the object's final velocity, and
- the resultant force acting on the object

all in the correct units!

Remember to check the units!

Step 2 : Decide how to tackle the problem

What is being asked? We are asked to calculate the time taken Δt to accelerate the object from the given initial velocity to final velocity. From the Law of Momentum,

$$\begin{aligned}\overrightarrow{F}_{Res}\Delta t &= \Delta\overrightarrow{p}\\ &= m\overrightarrow{v} - m\overrightarrow{u}\\ &= m(\overrightarrow{v} - \overrightarrow{u}).\end{aligned}$$

Thus we have everything we need to find Δt!

Step 3 : Choose a positive direction

Although not explicitly stated, the resultant force acts to the right. This follows from the fact that the object's velocity increases in this direction. Let us then choose right as the positive direction.

Step 4 : Do the calculation and quote the final answer

Right is the positive direction

$$\begin{aligned}\overrightarrow{F}_{Res}\Delta t &= m(\overrightarrow{v} - \overrightarrow{u})\\ (+150N)\Delta t &= (300kg)((+6\frac{m}{s}) - (+2\frac{m}{s}))\\ (+150N)\Delta t &= (300kg)(+4\frac{m}{s})\\ \Delta t &= \frac{(300kg)(+4\frac{m}{s})}{+150N}\\ \Delta t &= 8s\end{aligned}$$

Worked Example 38

Calculating Impulse

Question: A cricket ball weighing 156g is moving at 54 $km.hr^{-1}$ towards a batsman. It is hit by the batsman back towards the bowler at 36 $km.hr^{-1}$. Calculate i) the ball's impulse, and ii) the average force exerted by the bat if the ball is in contact with the bat for $0.13s$.

Answer:

Step 1 : Decide what information is supplied

The question explicitly gives

- the ball's mass,
- the ball's initial velocity,
- the ball's final velocity, and
- the time of contact between bat and ball

all except the time in the wrong units!

Remember to check the units!

Answer to (i):

Step 2 : Decide how to tackle the problem

What is being asked? We are asked to calculate the impulse

$$\text{Impulse} = \Delta\overrightarrow{p} = \overrightarrow{F}_{Res}\Delta t.$$

Since we do not have the force exerted by the bat on the ball ($\overrightarrow{F}_{Res}$), we have to calculate the impulse from the change in momentum of the ball. Now, since

$$\begin{aligned}\Delta\overrightarrow{p} &= \overrightarrow{p}_{final} - \overrightarrow{p}_{initial}\\ &= m\overrightarrow{v} - m\overrightarrow{u},\end{aligned}$$

we need the ball's mass, initial velocity and final velocity, which we are given.

Step 3 : Convert to S.I. units

Firstly let us change units for the mass

$$\begin{aligned}1000g &= 1kg\\ 1 &= \frac{1kg}{1000g}\\ 156g \times 1 &= 156g \times \frac{1kg}{1000g}\\ &= 0.156kg\end{aligned}$$

Next we change units for the velocity

$$\begin{aligned}1km &= 1000m\\ 1 &= \frac{1000m}{1km}\end{aligned}$$

$$\begin{aligned}3600s &= 1hr\\ 1 &= \frac{1hr}{3600s}\end{aligned}$$

$$\begin{aligned}54\frac{km}{hr} \times 1 \times 1 &= 54\frac{km}{hr} \times \frac{1000m}{1km} \times \frac{1hr}{3600s}\\ &= 15\frac{m}{s}\end{aligned}$$

$$
\begin{aligned}
36\frac{km}{hr} \times 1 \times 1 &= 36\frac{km}{hr} \times \frac{1000m}{1km} \times \frac{1hr}{3600s} \\
&= 10\frac{m}{s}
\end{aligned}
$$

Step 4 : Choose your convention
Next we must choose a positive direction. Let us choose the direction from the batsman to the bowler as the postive direction. Then the initial velocity of the ball is $\overrightarrow{u} = -15\ m.s^{-1}$, while the final velocity of the ball is $\overrightarrow{v} = +10\ m.s^{-1}$
Step 5 : Calculate the momentum
Now we calculate the change in momentum,

Direction from batsman to bowler is the positive direction

$$
\begin{aligned}
\Delta\overrightarrow{p} &= \overrightarrow{p}_{final} - \overrightarrow{p}_{initial} \\
&= m\overrightarrow{v} - m\overrightarrow{u} \\
&= m(\overrightarrow{v} - \overrightarrow{u}) \\
&= (0.156kg)((+10\ m.s^{-1}) - (-15\ m.s^{-1})) \\
&= +3.9\ kg.m.s^{-1} \\
&= 3.9\ kg.m.s^{-1}\ \textbf{in the direction from batsman to bowler}
\end{aligned}
$$

where we remembered in the last step to include the direction of the change in momentum in words.
Step 6 : Determine the impulse
Finally since impulse is just the change in momentum of the ball,

$$
\begin{aligned}
\text{Impulse} &= \Delta\overrightarrow{p} \\
&= 3.9\ kg.m.s^{-1}\ \textbf{in the direction from batsman to bowler}
\end{aligned}
$$

Answer to (ii):
Step 7 : Determine what is being asked
What is being asked? We are asked to calculate the average force exerted by the bat on the ball, $\overrightarrow{F}_{Res}$. Now,

$$\text{Impulse} = \overrightarrow{F}_{Res}\Delta t = \Delta\overrightarrow{p}.$$

We are given Δt and we have calculated the change in momentum or impulse of the ball in part (i)!
Step 8 : Choose a convention
Next we choose a positive direction. Let us choose the direction from the batsman to the bowler as the postive direction.
Step 9 : Calculate the force
Then substituting,

Direction from batsman to bowler is the positive direction

$$
\begin{aligned}
\overrightarrow{F}_{Res}\Delta t &= \text{Impulse} \\
\overrightarrow{F}_{Res}(0.13s) &= +3.9\frac{kg.m}{s} \\
\overrightarrow{F}_{Res} &= \frac{+3.9\frac{kg.m}{s}}{0.13s} \\
&= +30\frac{kg.m}{s^2} \\
&= 30N \textbf{ in the direction from batsman to bowler}
\end{aligned}
$$

where we remembered in the final step to include the direction of the force in words.

7.6 Summary of Important Quantities, Equations and Concepts

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
Momentum	$\overrightarrow{p}$	-	$kg.m.s^{-1}$	✓
Mass	m	-	kg	—
Velocity	$\overrightarrow{u},\overrightarrow{v}$	-	$m.s^{-1}$	✓
Change in momentum	$\Delta\overrightarrow{p}$	-	$kg.m.s^{-1}$	✓
Force	$\overrightarrow{F}$	N	$kg.m.s^{-2}$	✓
Impulse	J	-	$kg.m.s^{-1}$	✓

Table 7.1: Summary of the symbols and units of the quantities used in **Momentum**

Momentum The *momentum* of an object is defined as its mass multiplied by its velocity.

Momentum of a System The *total momentum of a system* is the sum of the momenta of each of the objects in the system.

Principle of Conservation of Linear Momentum: 'The total linear momentum of an isolated system is constant' or 'In an isolated system the total momentum before a collision (or explosion) is equal to the total momentum after the collision (or explosion)'.

Law of Momentum: The applied resultant force acting on an object is equal to the rate of change of the object's momentum and this force is in the direction of the change in momentum.

Chapter 8

Work and Energy

8.1 What are Work and Energy?

During this chapter you will discover that work and energy are very closely related: We consider the *energy* of an object as its *capacity to do work* and doing *work* as the process of transferring energy from one object or form to another. In other words,

- an object with lots of energy can do lots of work.
- when work is done, energy is lost by the object doing work and gained by the object on which the work is done.

Lifting objects or throwing them requires that you do work on them. Even making electricity flow requires that something do work. Something must have energy and transfer it through doing work to make things happen.

8.2 Work

To do work on an object, one must *move* the object by applying a *force* with at least a component in the direction of motion. The work done is given by

$$W = F_{||}s$$

W	: work done ($N.m$ or J)
$F_{||}$	: component of applied force parallel to motion (N)
s	: displacement of the object (m)

It is very important to note that for work to be done there must be a component of the applied force in the direction of motion. Forces perpendicular to the direction of motion do no work.

As with all physical quantities, work must have units. As follows from the definition, work is measured in $N.m$. The name given to this combination of S.I. units is the joule (J).

Definition: 1 joule is the work done when an object is moved $1m$ under the application of a force of $1N$ in the direction of motion.

The work done by an object can be positive or negative. Since force ($F_{||}$) and displacement (s) are both vectors, the result of the above equation depends on their directions:

- If $F_{||}$ acts in the same direction as the motion then positive work is being done. In this case the object on which the force is applied gains energy.
- If the direction of motion and $F_{||}$ are opposite, then negative work is being done. This means that energy is transferred in the opposite direction. For example, if you try to push a car uphill by applying a force up the slope and instead the car rolls down the hill you are doing negative work on the car. Alternatively, the car is doing positive work on you!

Worked Example 39

Calculating Work Done I

Question: If you push a box $20m$ forward by applying a force of $15N$ in the forward direction, what is the work you have done on the box?
Answer:
Step 1 : Analyse the question to determine what information is provided

- The force applied is $F = 15N$.
- The distance moved is $s = 20m$.
- The applied force and distance moved are in the same direction. Therefore, $F_{||} = 15N$.

These quantities are all in the correct units, so no unit conversions are required.

Step 2 : Analyse the question to determine what is being asked

- We are asked to find the work done on the box. We know from the definition that work done is $W = F_{||}s$

Step 3 : Next we substitute the values and calculate the work done

$$\begin{aligned} W &= F_{||}s \\ &= (15N)(20m) \\ &= 300\ N \cdot m \\ &= 300\ J \end{aligned}$$

Remember that the answer must be *positive* as the applied force and the motion are in the same direction (forwards). In this case, you (the pusher) lose energy, while the box gains energy.

Worked Example 40

Calculating Work Done II

Question: What is the work done by you on a car, if you try to push the car up a hill by applying a force of $40N$ directed up the slope, but it slides downhill $30cm$?

Answer:

Step 1 : Analyse the question to determine what information is provided

- The force applied is $F = 40N$
- The distance moved is $s = 30cm$. This is expressed in the wrong units so we must convert to the proper S.I. units (meters):

$$s = 30\text{cm} = 30\text{cm} \cdot \frac{1\text{m}}{100\text{cm}} = 0.3\text{m}$$

- The applied force and distance moved are in opposite directions. Therefore, if we take $s = 0.3m$, then $F_{||} = -40N$.

Step 2 : Analyse the question to determine what is being asked

- We are asked to find the work done on the car by you. We know that work done is $W = F_{||}s$

Step 3 : Substitute the values and calculate the work done

Again we have the applied force and the distance moved so we can proceed with calculating the work done:

$$\begin{aligned} W &= F_{||}s \\ &= (-40N)(0.3m) \\ &= -12N \cdot m \\ &= -12\ J \end{aligned}$$

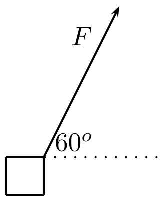
Note that the answer must be *negative* as the applied force and the motion are in opposite directions. In this case the car does work on the person trying to push.

What happens when the applied force and the motion are not parallel? If there is an angle between the direction of motion and the applied force then to determine the work done we have to calculate the *component* of the applied force *parallel* to the direction of motion. Note that this means a force perpendicular to the direction of motion can do no work.

Worked Example 41

Calculating Work Done III

Question: Calculate the work done on a box, if it is pulled $5m$ along the ground by applying a force of $F = 10N$ at an angle of 60^o to the horizontal.



Answer:

Step 1 : Analyse the question to determine what information is provided

- The force applied is $F = 10N$
- The distance moved is $s = 5m$ along the ground
- The angle between the applied force and the motion is 60^o

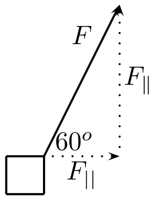
These quantities are in the correct units so we do not need to perform any unit conversions.

Step 2 : Analyse the question to determine what is being asked

- We are asked to find the work done on the box.

Step 3 : Calculate the component of the applied force in the direction of motion

Since the force and the motion are not in the same direction, we must first calculate the component of the force in the direction of the motion.



From the force diagram we see that the component of the applied force parallel to the ground is

$$\begin{aligned} F_{||} &= F \cdot \cos(60^o) \\ &= 10N \cdot \cos(60^o) \\ &= 5\ N \end{aligned}$$

Step 4 : Substitute and calculate the work done

Now we can calculate the work done on the box:

$$\begin{aligned} W &= F_{||} s \\ &= (5N)(5m) \\ &= 25\ J \end{aligned}$$

Note that the answer is positive as the component of the force $F_{||}$ is in the same direction as the motion.

We will now discuss energy in greater detail.

8.3 Energy

As we mentioned earlier, energy is the capacity to do work. When positive work is done on an object, the system doing the work loses energy. In fact, **the energy lost by a system is exactly equal to the work done by the system.**

Like work (W) the unit of energy (E) is the joule (J). This follows as work is just the transfer of energy.

A very important property of our universe which was discovered around 1890 is that energy is conserved.

Energy is never created nor destroyed, but merely transformed from one form to another.

Energy is **conserved!**

Energy conservation and the conservation of matter are the principles on which classical mechanics is built.

IN THE ABSENCE OF FRICTION

When work is done on an object by a system:

-the object gains energy equal to the work done by the system

Work Done = Energy Transferred

IN THE PRESENCE OF FRICTION

When work is done by a system:

-only some of the energy lost by the system is transferred into useful energy
-the rest of the energy transferred is lost to friction

Total Work Done = Useful Work Done + Work Done Against Friction

8.3.1 Types of Energy

So what different types of energy exist? Kinetic, mechanical, thermal, chemical, electrical, radiant, and atomic energy are just some of the types that exist. By the principle of conservation of energy, when work is done energy is merely transferred from one object to another and from one type of energy to another.

Kinetic Energy

Kinetic energy is the energy of motion that an object has. Objects moving in straight lines possess *translational kinetic energy*, which we often abbreviate as E_k.

The translational kinetic energy of an object is given by

$$\boxed{\begin{array}{rl} \multicolumn{2}{c}{E_k = \frac{1}{2}mv^2} \\ \\ E_k & : \text{kinetic energy } (J) \\ m & : \text{mass of object } (kg) \\ v & : \text{speed of the object } (m.s^{-1}) \end{array}}$$

Note the dependence of the kinetic energy on the speed of the object– kinetic energy is related to motion. The faster an object is moving the greater its kinetic energy.

Worked Example 42

Calculation of Kinetic Energy

Question: If a rock has a mass of $1kg$ and is thrown at $5m/s$, what is its kinetic energy?

Answer:

Step 1 : Analyse the question to determine what information is provided

- The mass of the rock $m = 1kg$
- The speed of the rock $v = 5m/s$

These are both in the correct units so we do not have to worry about unit conversions.

Step 2 : Analyse the question to determine what is being asked

- We are asked to find the kinetic energy. From the definition we know that to work out E_k, we need to know the mass and the velocity of the object and we are given both of these values.

Step 3 : Substitute and calculate the kinetic energy

$$\begin{aligned} E_k &= \frac{1}{2}mv^2 \\ &= \frac{1}{2}(1kg)(5\frac{m}{s})^2 \\ &= 12.5\frac{kg \cdot m^2}{s^2} \\ &= 12.5\ J \end{aligned}$$

To check that the units in the above example are in fact correct:

$$\begin{aligned}\frac{kg \cdot m^2}{s^2} &= \left(\frac{kg \cdot m}{s^2}\right) \cdot m = N \cdot m \\ &= J\end{aligned}$$

The units are indeed correct!

Study hint: Checking units is an important cross-check and you should get into a habit of doing this. If you, for example, finish an exam early then checking the units in your calculations is a very good idea.

Worked Example 43

Mixing Units and Kinetic Energy Calculations 1

Question: If a car has a mass of $900kg$ and is driving at $60km/hr$, what is its kinetic energy?

Answer:

Step 1 : Analyse the question to determine what information is provided

- The mass of the car $m = 900kg$
- The speed of the car $v = 60km/hr$. These are not the units we want so before we continue we must convert to m/s. We do this by *multiplying by one*:

$$\begin{aligned}60\frac{km}{hr} \times 1 &= 60\frac{km}{hr} \times \frac{1000m}{1km} \\ &= 60000\frac{m}{hr}\end{aligned}$$

Now we need to change from hours to seconds so we repeat our procedure:

$$\begin{aligned}60000\frac{m}{hr} \times 1 &= 60000\frac{m}{hr} \times \frac{1hr}{3600s} \\ &= 16.67\frac{m}{s}\end{aligned}$$

and so the speed in the units we want is $v = 16.67m/s$.

Step 2 : Analyse the question to determine what is being asked

- We are asked to find the kinetic energy.

Step 3 : Substitute and calculate

We know we need the mass and the speed to work out E_k and we are given both of these quantities. We thus simply substitute them into the equation for E_k:

$$\begin{aligned} E_k &= \frac{1}{2}mv^2 \\ &= \frac{1}{2}(900kg)(16.67\frac{m}{s})^2 \\ &= 125\ 000\frac{kgm^2}{s^2} \\ &= 125\ 000\ J \end{aligned}$$

Worked Example 44

Mixing Units and Kinetic Energy Calculations 2

Question: If a bullet has a mass of $150g$ and is shot at a muzzle velocity of $960m/s$, what is its kinetic energy?

Answer:

Step 1 : Analyse the question to determine what information is provided

- We are given the mass of the bullet $m = 150g$. This is not the unit we want mass to be in. We need to convert to kg. Again, we *multiply by one*:

$$\begin{aligned} 150g \cdot 1 &= 150g \cdot \frac{1kg}{1000g} \\ &= 0.15kg \end{aligned}$$

- We are given the muzzle velocity which is just how fast the bullet leaves the barrel and it is $v = 960m/s$.

Step 2 : Analyse the question to determine what is being asked

- We are asked to find the kinetic energy.

Step 3 : Substitute and calculate

We just substitute the mass and velocity (which are known) into the equation for E_k:

$$\begin{aligned} E_k &= \frac{1}{2}mv^2 \\ &= \frac{1}{2}(150kg)(960\frac{m}{s})^2 \\ &= 69\ 120\frac{kgm^2}{s^2} \\ &= 69\ 120\ J \end{aligned}$$

Potential Energy

If you lift an object you have to do work on it. This means that energy is transferred to the object. But where is this energy? This energy is stored in the object and is called *potential energy*. The reason it is called potential energy is because if we let go of the object it would move.

Definition: Potential energy is the energy an object has due to its position or state.

As an object raised above the ground falls, its *potential energy* is released and transformed into kinetic energy. The further it falls the faster it moves as more of the stored potential energy is transferred into kinetic energy. Remember, energy is never created nor destroyed, but merely transformed from one type to another. In this case potential energy is lost but an equal amount of kinetic energy is gained.

In the example of a falling mass the potential energy is known as *gravitational potential energy* as it is the gravitational force exerted by the earth which causes the mass to accelerate towards the ground. The gravitational field of the earth is what does the work in this case.

Another example is a rubber-band. In order to stretch a rubber-band we have to do work on it. This means we transfer energy to the rubber-band and it gains potential energy. This potential energy is called *elastic potential energy*. Once released, the rubber-band begins to move and elastic potential energy is transferred into kinetic energy.

Gravitational Potential Energy

As we have mentioned, when lifting an object it gains gravitational potential energy. One is free to define any level as corresponding to zero gravitational potential energy. Objects above this level then possess positive potential energy, while those below it have negative potential energy. To avoid negative numbers in a problem, always choose the lowest level as the zero potential mark. The change in gravitational potential energy of an object is given by:

$$\Delta E_P = mg\Delta h$$

ΔE_P	: Change in gravitational potential energy (J)
m	: mass of object (kg)
g	: acceleration due to gravity $(m.s^{-2})$
Δh	: change in height (m)

When an object is lifted it gains gravitational potential energy, while it loses gravitational potential energy as it falls.

Worked Example 45

Gravitational potential energy

Question: How much potential energy does a brick with a mass of $1kg$ gain if it is lifted $4m$.
Answer:
Step 1 : Analyse the question to determine what information is provided

- The mass of the brick is $m = 1kg$
- The height lifted is $\Delta h = 4m$

These are in the correct units so we do not have to worry about unit conversions.
Step 2 : Analyse the question to determine what is being asked

- We are asked to find the gain in potential energy of the object.

Step 3 : Identify the type of potential energy involved
Since the block is being lifted we are dealing with gravitational potential energy. To work out ΔE_P, we need to know the mass of the object and the height lifted. As both of these are given, we just substitute them into the equation for ΔE_P.
Step 4 : Substitute and calculate

$$\begin{aligned}\Delta E_P &= mg\Delta h \\ &= (1kg)\left(10\frac{m}{s^2}\right)(4m) \\ &= 40\frac{kg \cdot m^2}{s^2} \\ &= 40\ J\end{aligned}$$

8.4 Mechanical Energy and Energy Conservation

Kinetic energy and potential energy are together referred to as mechanical energy. The total mechanical energy (U) of an object is then the sum of its kinetic and potential energies:

$$\begin{aligned}U &= E_P + E_K \\ U &= mgh + \frac{1}{2}mv^2 \qquad (8.1)\end{aligned}$$

Now,

IN THE ABSENCE OF FRICTION

Mechanical energy is conserved

$$U_{before} = U_{after}$$

This principle of conservation of mechanical energy can be a very powerful tool for solving physics problems. However, in the presence of friction some of the mechanical energy is lost:

IN THE PRESENCE OF FRICTION

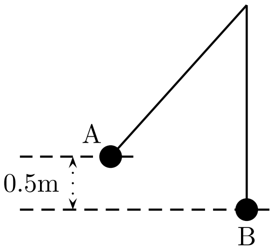
Mechanical energy is **not** conserved
(The mechanical energy lost is equal
to the work done against friction)

$$\Delta U = U_{before} - U_{after} = \text{Work Done Against Friction}$$

Worked Example 46

Using Mechanical Energy Conservation

Question: A $2kg$ metal ball is suspended from a rope. If it is released from point A and swings down to the point B (the bottom of its arc) what is its velocity at point B?



Answer:

Step 1 : Analyse the question to determine what information is provided

- The mass of the metal ball is $m = 2kg$
- The change in height going from point A to point B is $h = 0.5m$
- The ball is released from point A so the velocity at point A is zero ($v_A = 0m/s$).

These are in the correct units so we do not have to worry about unit conversions.

Step 2 : Analyse the question to determine what is being asked

- Find the velocity of the metal ball at point B.

Step 3 : Determine the Mechanical Energy at A and B

To solve this problem we use conservation of mechanical energy as there is no friction. Since mechanical energy is conserved,

$$U_A = U_B$$

Therefore we need to know the mechanical energy of the ball at point A (U_A) and at point B (U_B). The mechanical energy at point A is

$$U_A = mgh_A + \frac{1}{2}m(v_A)^2$$

We already know m, g and v_A, but what is h_A? Note that if we let $h_B = 0$ then $h_A = 0.5m$ as A is $0.5m$ above B. In problems you are always free to choose a line corresponding to $h = 0$. In this example the most obvious choice is to make point B correspond to $h = 0$.

Now we have,

$$\begin{aligned} U_A &= (2kg)\left(10\frac{m}{s^2}\right)(0.5m) + \frac{1}{2}(2kg)(0)^2 \\ &= 10\ J \end{aligned}$$

As already stated $U_B = U_A$. Therefore $U_B = 10J$, but using the definition of mechanical energy

$$\begin{aligned} U_B &= mgh_B + \frac{1}{2}m(v_B)^2 \\ &= \frac{1}{2}m(v_B)^2 \end{aligned}$$

because $h_B = 0$. This means that

$$\begin{aligned} 10J &= \frac{1}{2}(2kg)(v_B)^2 \\ (v_B)^2 &= 10\frac{J}{kg} \\ v_B &= \sqrt{10}\frac{m}{s} \end{aligned}$$

8.5 Summary of Important Quantities, Equations and Concepts

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
Work	W	J	$N.m$ **or** $kg.m^2.s^{-2}$	—
Kinetic Energy	E_K	J	$N.m$ **or** $kg.m^2.s^{-2}$	—
Potential Energy	E_P	J	$N.m$ **or** $kg.m^2.s^{-2}$	—
Mechanical Energy	U	J	$N.m$ **or** $kg.m^2.s^{-2}$	—

Table 8.1: Summary of the symbols and units of the quantities used in **Energy**

Principle of Conservation of Energy: Energy is never created nor destroyed, but merely transformed from one form to another.

Conservation of Mechanical Energy: In the absence of friction, the total mechanical energy of an object is conserved.

Essay 1: Energy

Author: Asogan Moodaly

Asogan Moodaly received his Bachelor of Science degree (with honours) in Mechanical Engineering from the University of Natal, Durban in South Africa. For his final year design project he worked on a 3-axis filament winding machine for composite (Glass re-enforced plastic in this case) piping. He worked in Vereeniging, Gauteng at Mine Support Products (a subsidiary of Dorbyl Heavy Engineering) as the design engineer once he graduated. He currently lives in the Vaal Triangle area and is working for Sasol Technology Engineering as a mechanical engineer, ensuring the safety and integrity of equipment installed during projects.

Energy and electricity. Why the fuss?

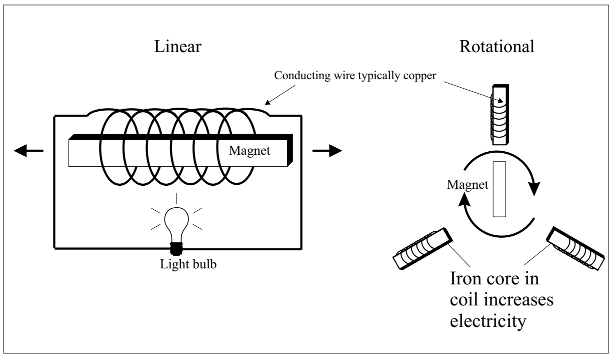
Disclaimer: The details furnished below are very basic and for illustration purposes only.

Why do we need energy? Note that I use the word energy and not electricity. On a broad scale it stimulates economic growth, etc, etc but on a personal level it allows us to lead a comfortable lifestyle.

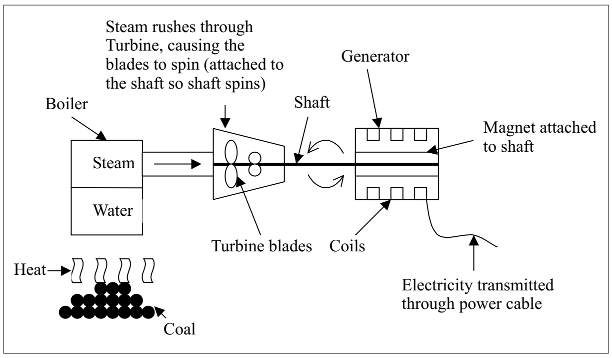
e.g. Flick a switch and Heat for cooking Entertainment such as television and radio Heat for water and interior of house Ironing Electronic and electrical devices such as alarms, garage doors, etc.

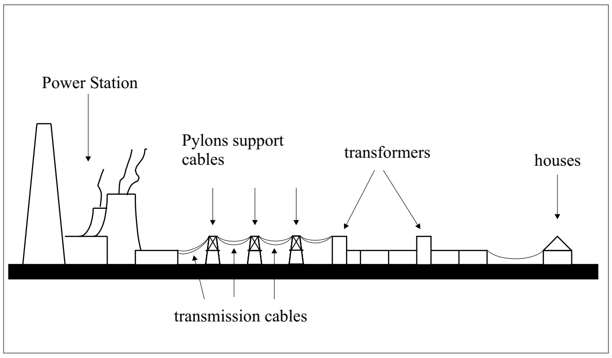
In a modern household this energy is provided in the form of electricity which is powered via fossil fuels or nuclear.

How is electricity made? In a nutshell: By moving a magnet through or near a set of conducting coils.


Linear
Rotational
Conducting wire typically copper
Magnet
Light bulb
Magnet
Iron core in coil increases electricity

Most power stations produce steam through heat (nuclear reaction or burning fossil fuels), the steam drives a turbine which moves a magnet relative to a coil (the generator like the above but on a much larger scale i.e. bigger magnets, bigger coils, etc), which produces electricity that is transmitted via a power network to our homes. Gas fired plants burn gas directly in a gas turbine to produce the same desired relative motion between permanent magnet and coil.


Steam rushes through Turbine, causing the blades to spin (attached to the shaft so shaft spins)
Generator
Boiler
Shaft
Magnet attached to shaft
Steam
Water
Turbine blades
Coils
Heat
Electricity transmitted through power cable
Coal



Coal, oil and gas are fossil fuels. Fossil fuels were created by decomposing organic (plant and animal) matter a long, long time ago and are typically found underground. Different temperatures and pressures resulted in the organic matter transforming into coal, oil or gas.

Why the fuss about fossil fuels?

1. Fossil fuel power is bad news in the long run. It pollutes and contributes to the greenhouse effect (global warming resulting in melting polar ice caps, floods, droughts, disease, etc).

2. Its not going to last forever.

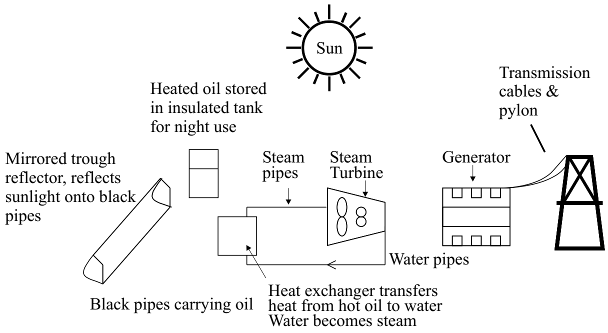
3. Nuclear power is cleaner in terms of emissions but theres no proven way of disposing of the nuclear waste. Oh, and it wont last forever either!

Renewable Energy As the name suggests renewable energy lasts forever. Solar (sun), wind, geothermal, wave, hydro and biomass (organic) are all sources of energy that will last until the sun eventually explodes many millions of years from now. Hopefully the human race will have moved from the earth by then! Generally the principal of renewable electricity generation is similar to fossil fuel electricity generation in that electricity is generated by moving a magnet relative to a conducting coil. What is different is the way energy is supplied to cause that motion.

The below are a few different types of available renewable energy technologies.

Solar

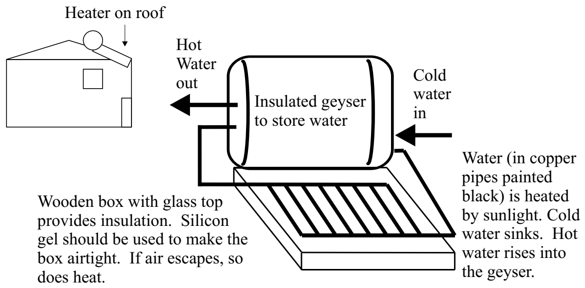
There are different types of solar electricity technologies, the main ones being solar thermal and photovoltaic.


Sun
Heated oil stored in insulated tank for night use
Transmission cables & pylon
Mirrored trough reflector, reflects sunlight onto black pipes
Steam pipes
Steam Turbine
Generator
Water pipes
Black pipes carrying oil
Heat exchanger transfers heat from hot oil to water
Water becomes steam

Solar thermal uses the heat of the sun to produce electricity. Sun is concentrated using mirrors. This heat either creates steam which drives a turbine which in turn drives a generator (as per fossil fuel generation), or drives an air engine (engine that uses expanding air to obtain motion) that drives a generator.

Photovoltaic panels convert sunlight directly into electricity. The benefit of photovoltaic panels is that there are no moving parts, and is therefore relatively maintenance free. The downside is that its very expensive at this stage (17/06/2004).

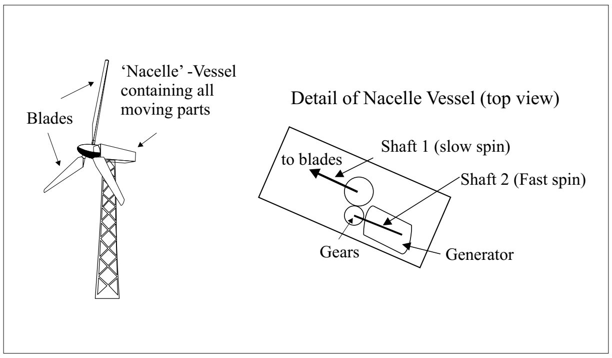
Solar Water Heaters could save up to 30% of the total electricity used in a house.


Heater on roof
Hot Water out
Insulated geyser to store water
Cold water in
Water (in copper pipes painted black) is heated by sunlight. Cold water sinks. Hot water rises into the geyser.
Wooden box with glass top provides insulation. Silicon gel should be used to make the box airtight. If air escapes, so does heat.

Wind

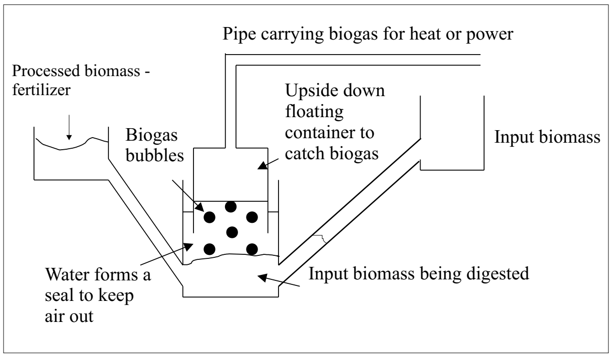
Wind turbines catch wind that spins the blades. The blades are connected to a shaft that spins because of the wind. This spinning shaft spins another shaft that turns a permanent magnet relative to conducting coils.

Note that gears are used to convert the slow spinning of the 1st shaft to a faster spin on the 2nd shaft. The generator shaft needs to spin at the correct speed to produce the right amount and quality of electricity. Some generators are now being modified to run at slower speeds. This saves money as gears are not needed.


'Nacelle' -Vessel containing all moving parts
Blades
Detail of Nacelle Vessel (top view)
to blades
Shaft 1 (slow spin)
Shaft 2 (Fast spin)
Gears
Generator

Biomass

Biomass is anything organic i.e. plant or animal matter. It can be used in the place of coal as per a normal coal fired plant and is renewable as long as the biomass e.g. wood; is handled in a sustainable manner. By sustainable I mean that suitable farming practices are used so that the land is not over farmed which will result in the soil becoming barren and nothing growing there again.


Pipe carrying biogas for heat or power
Processed biomass - fertilizer
Biogas bubbles
Upside down floating container to catch biogas
Input biomass
Water forms a seal to keep air out
Input biomass being digested

Biomass can also be processed using anaerobic digestion to produce a gas that can be burned for heat or electricity. This biogas is made up of a number of other gases that are similar to those found in fossil fuel natural gas Except the amount of the gases are different. E.g. Natural gas has about 94

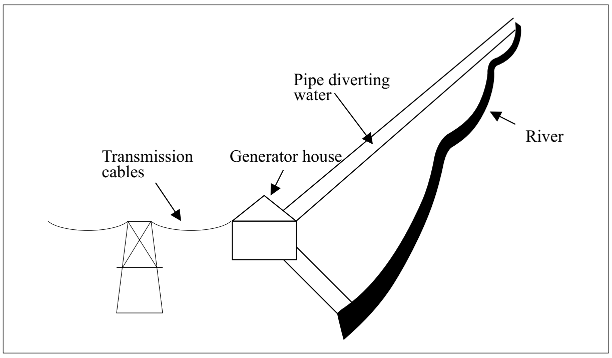
Anaerobic digestion: Anaerobic means No air. Therefore anaerobic digestion means to digest in the absence of air. Bacteria that naturally exist in organic matter will convert organic matter to biogas and fertilizer when all the air is removed.

Thousands of anaerobic digesters have been installed in rural India, Nepal and China in rural areas where cow dung, human waste and chicken litter (faeces) are all processed using anaerobic digestion to produce gas that can be burned in the home for cooking and heating. The leftover is used as fertilizer.

Geothermal Energy

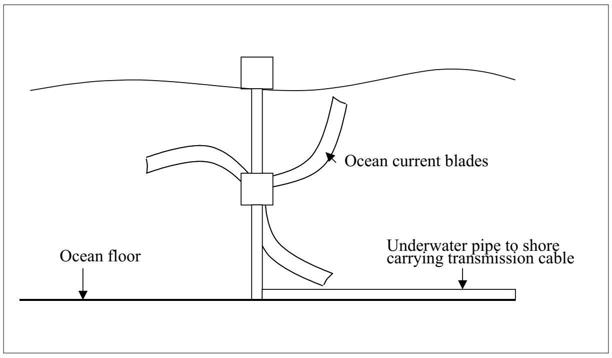
In some places on earth, the earths crust is thinner than others. As a result the heat from the earths core escapes. The heat can be captured by converting water to steam, and using the steam to drive a steam generator as discussed above.

Hydroelectric power Water from a river is diverted to turn a water turbine to create electricity similar to the principles of steam generation. The water is returned to the river after driving the turbine.

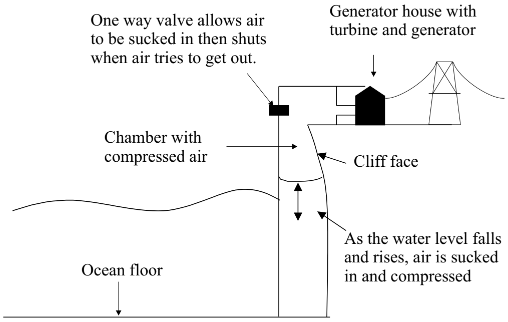


Wave Energy

Some wave energy generators work similarly to wind turbines except that underwater ocean currents turns the blades instead of wind; and of course most of the structure is under water!



Another concept uses the rising and falling of the tides to suck air in using a one way valve. As a result air becomes compressed in a chamber and the compressed air is let out to drive a turbine which in turn drives a generator


One way valve allows air to be sucked in then shuts when air tries to get out.
Generator house with turbine and generator
Chamber with compressed air
Cliff face
As the water level falls and rises, air is sucked in and compressed
Ocean floor

These are relatively new technologies.

Liquid Fuels Liquid fuels are used mainly for transportation. Petrol and diesel are the most common liquid fuels and are obtained from oil.

Sasol is the only company in the world that makes liquid fuels from coal; and will be one of the leading companies in the world to make liquid fuels from natural gas! The Sasol petro-chemical plants are based in Sasolburg on the border of the Free State and in Secunda in Mpumalanga.

However, as discussed above coal, gas and oil are fossil fuels and are not renewable. Petrol and diesel are obtained from fossil fuels and therefore pollute and contribute to the green house effect (global warming).

Alternatives

Biodiesel

Oil can be extracted from plants such as the soya bean, sunflower and rapeseed by pressing it through a filter. This oil if mixed correctly with either methanol or dry ethanol and Sodium Hydroxide will separate the plant oil into biodiesel, glycerol and fertilizer.

The biodiesel can be used as produced in a conventional diesel engine with little or no modifications required.

The glycerol can be refined a bit further for pharmaceutical companies to use, or can be used to make soap.

Ethanol

Corn, maize and sugar cane can be used to make ethanol as a fuel substitute for petrol. Its made by the same fermentation process used to make alcohol. Enzymes are often used to speed up the process.

In ethanol from sugar cane production, the leftover bagasse (the fibre part of the sugar cane) can be burned in a biomass power station to produce electricity.

Hydrogen

Through the process of electrolysis electricity (hopefully clean, renewable electricity!) can split water into hydrogen and oxygen. The stored hydrogen can be used in a fuel cell to create electricity in a process that is opposite to electrolysis; to drive electric motors in a car.

The hydrogen can also be burned directly in a modified internal combustion engine.

In both cases the waste product is water.

Essay 2: Tiny, Violent Collisions

Author: Thomas D. Gutierrez

Tom Gutierrez received his Bachelor of Science and Master degrees in physics from San Jose State University in his home town of San Jose, California. As a Master's student he helped work on a laser spectrometer at NASA Ames Research Centre. The instrument measured the ratio of different isotopes of carbon in CO_2 gas and could be used for such diverse applications as medical diagnostics and space exploration. Later, he received his Ph.D. in physics from the University of California, Davis where he performed calculations for various reactions in high energy physics collisions. He currently lives in Berkeley, California where he studies proton-proton collisions seen at the STAR experiment at Brookhaven National Laboratory on Long Island, New York.

High Energy Collisions

Take an orange and expanded it to the size of the earth. The atoms of the earth-sized orange would themselves be about the size of regular oranges and would fill the entire "earth-orange". Now, take an atom and expand it to the size of a football field. The nucleus of that atom would be about the size of a tiny seed in the middle of the field. From this analogy, you can see that atomic nuclei are very small objects by human standards. They are roughly 10^{-15} meters in diameter – one-hundred thousand times smaller than a typical atom. These nuclei cannot be seen or studied via any conventional means such as the naked eye or microscopes. So how do scientists study the structure of very small objects like atomic nuclei?

The simplest nucleus, that of hydrogen, is called the proton. Faced with the inability to isolate a single proton, open it up, and directly examine what is inside, scientists must resort to a brute-force and somewhat indirect means of exploration: high energy collisions. By colliding protons with other particles (such as other protons or electrons) at very high energies, one hopes to learn about what they are made of and how they work. The American physicist Richard Feynman once compared this process to slamming delecate watches together and figuring out how they work by only examining the broken debris. While this analogy may seem pessimistic, with sufficent mathematical models and experimental precision, considerable information can be extracted from the debris of such high energy subatomic collisions. One can learn about both the nature of the forces at work and also about the sub-structure of such systems.

The experiments are in the category of "high energy physics" (also known as "subatomic" physics). The primary tool of scientific exploration in these experiments is an extremely violent collision between two very, very small subatomic objects such as nuclei. As a general rule, the higher the energy of the collisions, the more detail of the original system you are able to resolve. These experiments are operated at laboratories such as CERN, SLAC, BNL, and Fermilab, just to name a few. The giant machines that perform the collisions are roughly the size of towns. For example, the RHIC collider at BNL is a ring about 1 km in diameter and can be seen from space. The newest machine currently being built, the LHC at CERN, is a ring 9 km in diameter!

Let's examine the kinematics of such a collisions in some detail...

Chapter 9

Collisions and Explosions

In most physics courses questions about collisions and explosions occur and to solve these we must use the ideas of momentum and energy; with a bit of mathematics of course! This section allows you to pull the momentum and energy ideas together easily with some specific problems.

9.1 Types of Collisions

We will consider two types of collisions in this section

- Elastic collisions
- Inelastic collisions

In both types of collision, total energy and total momentum is *always* conserved. Kinetic energy is conserved for elastic collisions, but not for inelastic collisions.

9.1.1 Elastic Collisions

Definition: An *elastic collision* is a collision where total momentum and total kinetic energy are both conserved.
(NOTE TO SELF: this should be in an environment for definitions!!)

This means that the total momentum *and* the total kinetic energy before an elastic collision is the same as after the collision. For these kinds of collisions, the kinetic energy is not changed into another type of energy.

Before the Collision

In the following diagram, two balls are rolling toward each other, about to collide

$\overrightarrow{p}_1$, K_1
$\overrightarrow{p}_2$, K_2

Before the balls collide, the total momentum of the system is equal to all the individual momenta added together. The ball on the left has a momentum which we call $\overrightarrow{p}_1$ and the ball on the right has a momentum which we call $\overrightarrow{p}_2$, it means the total momentum before the collision is

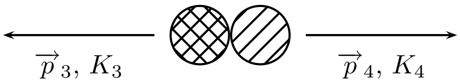
$$\overrightarrow{p}_{\text{Before}} = \overrightarrow{p}_1 + \overrightarrow{p}_2 \tag{9.1}$$

We calculate the total kinetic energy of the system in the same way. The ball on the left has a kinetic energy which we call K_1 and the ball on the right has a kinetic energy which we call K_2, it means that the total kinetic energy before the collision is

$$K_{\text{Before}} = K_1 + K_2 \tag{9.2}$$

After the Collision

The following diagram shows the balls after they collide



After the balls collide and bounce off each other, they have new momenta and new kinetic energies. Like before, the total momentum of the system is equal to all the individual momenta added together. The ball on the left now has a momentum which we call $\overrightarrow{p}_3$ and the ball on the right now has a momentum which we call $\overrightarrow{p}_4$, it means the total momentum after the collision is

$$\overrightarrow{p}_{\text{After}} = \overrightarrow{p}_3 + \overrightarrow{p}_4 \tag{9.3}$$

The ball on the left now has a kinetic energy which we call K_3 and the ball on the right now has a kinetic energy which we call K_4, it means that the total kinetic energy after the collision is

$$K_{\text{After}} = K_3 + K_4 \tag{9.4}$$

Since this is an *elastic* collision, the total momentum before the collision equals the total momentum after the collision **and** the total kinetic energy before the collision equals the total kinetic energy after the collision

$$\begin{array}{rcl} \text{Before} & & \text{After} \\ \overrightarrow{p}_{\text{Before}} & = & \overrightarrow{p}_{\text{After}} \\ \overrightarrow{p}_1 + \overrightarrow{p}_2 & = & \overrightarrow{p}_3 + \overrightarrow{p}_4 \end{array} \tag{9.5}$$

$$\textbf{and}$$

$$\begin{array}{rcl} K_{\text{Before}} & = & K_{\text{After}} \\ K_1 + K_2 & = & K_3 + K_4 \end{array} \tag{9.6}$$

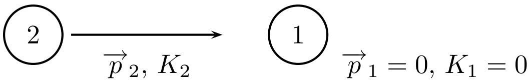
Worked Example 47

An Elastic Collision

We will have a look at the collision between two pool balls. Ball 1 is at rest and ball 2 is moving towards it with a speed of 2 [m.s^{-1}]. The mass of each ball is 0.3 [Kg]. After the balls collide *elastically*, ball 2 comes to a stop and ball 1 moves off. What is the final velocity of ball 1?

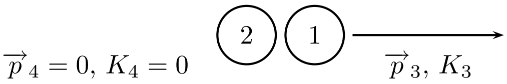
Step 1 : Draw the "before" diagram

Before the collision, ball 2 is moving; we will call it's momentum P_2 and it's kinetic energy K_2. Ball 1 is at rest, so it has zero kinetic energy and momentum.


2
$\overrightarrow{p}_2$, K_2
1
$\overrightarrow{p}_1 = 0$, $K_1 = 0$

Step 2 : Draw the "after" diagram

After the collision, ball 2 is at rest but ball 1 has a momentum which we call P_3 and a kinetic energy which we call K_3.


$\overrightarrow{p}_4 = 0$, $K_4 = 0$
2
1
$\overrightarrow{p}_3$, K_3

Because the collision is elastic, we can solve the problem using momentum conservation *or* kinetic energy conservation. We will do it both ways to show that the answer is the same, whichever method you use.

Step 3 : Show the conservation of momentum

We start by writing down that the momentum before the collision $\overrightarrow{p}_{\text{Before}}$ is equal to the momentum after the collision $\overrightarrow{p}_{\text{After}}$

$$\begin{aligned} \text{Before} & \quad \text{After} \\ \overrightarrow{p}_{\text{Before}} &= \overrightarrow{p}_{\text{After}} \\ \overrightarrow{p}_1 + \overrightarrow{p}_2 &= \overrightarrow{p}_3 + \overrightarrow{p}_4 \\ 0 + \overrightarrow{p}_2 &= \overrightarrow{p}_3 + 0 \\ \overrightarrow{p}_2 &= \overrightarrow{p}_3 \end{aligned} \tag{9.7}$$

We know that momentum is just $P = mv$, and we know the masses of the balls, so we can rewrite the conservation of momentum in terms of the velocities of the balls

$$\begin{aligned} \overrightarrow{p}_2 &= \overrightarrow{p}_3 \\ m_2 v_2 &= m_3 v_3 \\ 0.3 v_2 &= 0.3 v_3 \\ v_2 &= v_3 \end{aligned} \tag{9.8}$$

So ball 1 exits with the velocity that ball 2 started with!

$$v_3 = 2[\text{m.s}^{-1}] \tag{9.9}$$

Step 4 : Show the conservation of kinetic energy
We start by writing down that the kinetic energy before the collision K_{Before} is equal to the kinetic energy after the collision K_{After}

$$\begin{aligned}\text{Before} & \quad \text{After} \\ K_{\text{Before}} &= K_{\text{After}} \\ K_1 + K_2 &= K_3 + K_4 \\ 0 + K_2 &= K_3 + 0 \\ K_2 &= K_3\end{aligned} \tag{9.10}$$

We know that kinetic energy is just $K = \frac{mv^2}{2}$, and we know the masses of the balls, so we can rewrite the conservation of kinetic energy in terms of the velocities of the balls

$$\begin{aligned}K_2 &= K_3 \\ \frac{m_2 v_2^2}{2} &= \frac{m_3 v_3^2}{2} \\ 0.15v_2^2 &= 0.15v_3^2 \\ v_2^2 &= v_3^2 \\ v_2 &= v_3\end{aligned} \tag{9.11}$$

So ball 1 exits with the velocity that ball 2 started with, which agrees with the answer we got when we used the conservation of momentum.

$$v_3 = 2[\text{m.s}^{-1}] \tag{9.12}$$

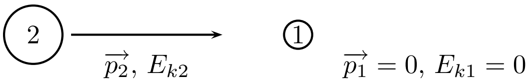
Worked Example 48

Elastic Collision 2

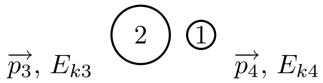
Question: Now for a slightly more difficult example. We have 2 marbles. Marble 1 has mass 50 g and marble 2 has mass 100 g. I roll marble 2 along the ground towards marble 1 in the positive x-direction. Marble 1 is initially at rest and marble 2 has a velocity of 3 m.s^{-1} in the positive x-direction. After they collide *elastically*, both marbles are moving. What is the final velocity of each marble?
Answer:
Step 1 : Put all the quantities into S.I. units
So:

$$m_1 = 0.05\text{kg} \quad \text{and} \quad m_2 = 0.1\text{kg}$$

Step 2 : Draw a rough sketch of the situation
<u>Before</u> the collision:


2
$\overrightarrow{p_2}$, E_{k2}
1
$\overrightarrow{p_1} = 0$, $E_{k1} = 0$

After the collision:


2
1
$\overrightarrow{p_3}$, E_{k3}
$\overrightarrow{p_4}$, E_{k4}

Step 3 : Decide which equations to use in the problem
Since the collision is *elastic*, both momentum *and* kinetic energy are conserved in the collision. So:

$$\begin{aligned} E_{kBefore} &= E_{kAfter} \\ &\text{and} \\ \overrightarrow{p_{Before}} &= \overrightarrow{p_{After}} \end{aligned}$$

There are two unknowns ($\overrightarrow{v_1}$ and $\overrightarrow{v_2}$) so we will need two equations to solve for them. We need to use both **kinetic energy conservation** and **momentum conservation** in this problem.
Step 4 : Solve the first equation
Let's start with energy conservation. Then:

$$\begin{aligned} E_{kBefore} &= E_{kAfter} \\ \frac{1}{2}m_1\overrightarrow{u_1}^2 + \frac{1}{2}m_2\overrightarrow{u_2}^2 &= \frac{1}{2}m_1\overrightarrow{v_1}^2 + \frac{1}{2}m_2\overrightarrow{v_2}^2 \\ m_1\overrightarrow{u_1}^2 + m_2\overrightarrow{u_2}^2 &= m_1\overrightarrow{v_1}^2 + m_2\overrightarrow{v_2}^2 \end{aligned}$$

But $\overrightarrow{u_1}$=0, and solving for $\overrightarrow{v_2}^2$:

$$\begin{aligned} \overrightarrow{v_2}^2 &= \overrightarrow{u_2}^2 - \frac{m1}{m2}\overrightarrow{v_1}^2 \\ \overrightarrow{v_2}^2 &= (3)^2 - \frac{(0.05)}{(0.10)}\overrightarrow{v_1}^2 \\ \overrightarrow{v_2}^2 &= 9 - \frac{1}{2}\overrightarrow{v_1}^2 \quad (\mathbf{A}) \end{aligned}$$

Step 5 : Solve the second equation
Now we have simplified as far as we can, we move onto momentum conservation:

$$\begin{aligned} \overrightarrow{p_{Before}} &= \overrightarrow{p_{After}} \\ m_1\overrightarrow{u_1} + m_2\overrightarrow{u_2} &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2} \end{aligned}$$

But $\overrightarrow{u_1}$=0, and solving for $\overrightarrow{v_1}$:

$$\begin{aligned} m_2\overrightarrow{u_2} &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2} \\ m_1\overrightarrow{v_1} &= m_2\overrightarrow{u_2} - m_2\overrightarrow{v_2} \\ \overrightarrow{v_1} &= \frac{m_2}{m_1}\overrightarrow{u_2} - \frac{m_2}{m_1}\overrightarrow{v_2} \\ \overrightarrow{v_1} &= 2(3) - 2\overrightarrow{v_2} \\ \overrightarrow{v_1} &= 6 - 2\overrightarrow{v_2} \quad (\mathbf{B}) \end{aligned}$$

Step 6 : Substitute one equation into the other

Now we can substitute (**B**) into (**A**) to solve for $\overrightarrow{v_2}$:

$$\begin{aligned}
\overrightarrow{v_2}^2 &= 9 - \frac{1}{2}\overrightarrow{v_1}^2 \\
\overrightarrow{v_2}^2 &= 9 - \frac{1}{2}(6 - 2\overrightarrow{v_2})^2 \\
\overrightarrow{v_2}^2 &= 9 - \frac{1}{2}(36 - 24\overrightarrow{v_2} + 4\overrightarrow{v_2}^2) \\
\overrightarrow{v_2}^2 &= 9 - 18 + 12\overrightarrow{v_2} - 2\overrightarrow{v_2}^2 \\
3\overrightarrow{v_2}^2 &= -9 + 12\overrightarrow{v_2} \\
\overrightarrow{v_2}^2 &= 4\overrightarrow{v_2} - 3 \\
\overrightarrow{v_2}^2 - 4\overrightarrow{v_2} + 3 &= 0 \\
(\overrightarrow{v_2} - 3)(\overrightarrow{v_2} - 1) &= 0 \\
\overrightarrow{v_2} = 3 \quad &\textbf{or} \quad \overrightarrow{v_2} = 1
\end{aligned}$$

We were lucky in this question because we could factorise. If you can't factorise, then you can always solve using the formula for solving quadratic equations. Remember:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

So, just to check:

$$\begin{aligned}
\overrightarrow{v_2} &= \frac{4 \pm \sqrt{4^2 - 4(1)(3)}}{2(1)} \\
\overrightarrow{v_2} &= \frac{4 \pm \sqrt{16 - 12}}{2} \\
\overrightarrow{v_2} &= \frac{4 \pm \sqrt{4}}{2} \\
\overrightarrow{v_2} &= 2 \pm 1 \\
\overrightarrow{v_2} = 3 \quad &\textbf{or} \quad \overrightarrow{v_2} = 1 \ \text{ same as before}
\end{aligned}$$

Step 7 : Solve for and quote the final answers

So finally, substituting into equation (**B**) to get $\overrightarrow{v_1}$:

$$\overrightarrow{v_1} = 6 - 2\overrightarrow{v_2}$$

If $\overrightarrow{v_2} = 3$ m.s^{-1} then

$$\overrightarrow{v_1} = 6 - 2(3) = 0 \text{ m.s}^{-1}$$

But, according to the question, marble 1 is *moving* after the collision. So $\overrightarrow{v_1} \neq 0$ and $\overrightarrow{v_2} \neq 3$. Therefore:

$$\begin{aligned}
\overrightarrow{v_2} &= \underline{1 \text{ m.s}^{-1} \quad \text{in the positive x} - \text{direction}} \\
&\text{and} \\
\overrightarrow{v_1} &= \underline{4 \text{ m.s}^{-1} \quad \text{in the positive x} - \text{direction}}
\end{aligned}$$

9.1.2 Inelastic Collisions

Definition: An *inelastic collision* is a collision in which **total momentum** is conserved but **total kinetic energy** is *not* conserved; the kinetic energy is *transformed* into other kinds of energy.

So the total momentum before an inelastic collisions is the same as after the collision. But the total *kinetic* energy before and after the inelastic collision is *different*. Of course this does not mean that total energy has not been conserved, rather the energy has been *transformed* into another type of energy.

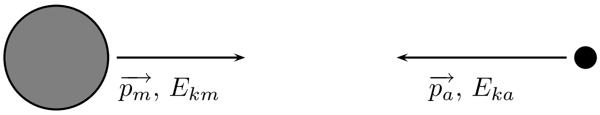
As a rule of thumb, inelastic collisions happen when the colliding objects are distorted in some way. Usually they change their shape. To modify the shape of an object requires energy and this is where the "missing" kinetic energy goes. A classic example of an inelastic collision is a car crash. The cars change shape and there is a noticeable change in the kinetic energy of the cars before and after the collision. This energy was used to bend the metal and deform the cars. Another example of an inelastic collision is shown in the following picture.

Here an asteroid (the small circle) is moving through space towards the moon (big circle). Before the moon and the asteroid collide, the total momentum of the system is:

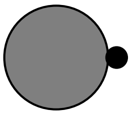
$$\overrightarrow{p_{Before}} = \overrightarrow{p_m} + \overrightarrow{p_a}$$

($\overrightarrow{p_m}$ stands for $\overrightarrow{p_{moon}}$ and $\overrightarrow{p_a}$ stands for $\overrightarrow{p_{asteroid}}$) and the total kinetic energy of the system is:

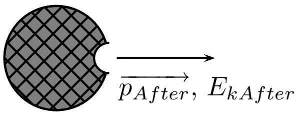
$$E_{Before} = E_{km} + E_{ka}$$


$\overrightarrow{p_m}$, E_{km}
$\overrightarrow{p_a}$, E_{ka}

When the asteroid collides **inelastically** with the moon, its kinetic energy is transformed mostly into heat energy. If this heat energy is large enough, it can cause the asteroid and the area of the moon's surface that it hit, to melt into liquid rock! From the force of impact of the asteroid, the molten rock flows outwards to form a moon crater.



After the collision, the total momentum of the system will be the same as before. But since this collision is *inelastic*, (and you can see that a change in the shape of objects has taken place!), total kinetic energy is **not** the same as before the collision.



So:

$$\begin{aligned}
\overrightarrow{p_{Before}} &= \overrightarrow{p_{After}} \\
\overrightarrow{p_m} + \overrightarrow{p_a} &= \overrightarrow{p_{After}} \\
&\textbf{but} \\
E_{kBefore} &\neq E_{kAfter} \\
E_{km} + E_{ka} &\neq E_{kAfter}
\end{aligned}$$

Worked Example 49

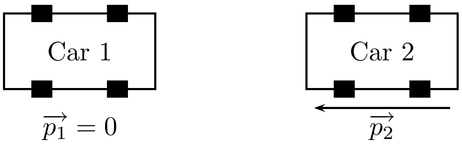
Inelastic Collision

Question: Let's consider the collision of two cars. Car 1 is at rest and Car 2 is moving at a speed of 2m.s^{-1} in the negative x-direction. Both cars each have a mass of 500kg. The cars collide *inelastically* and stick together. What is the resulting velocity of the resulting mass of metal?

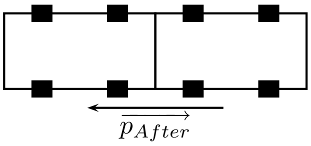
Answer:

Step 1 : Draw a rough sketch of the situation

Before the collision:



After the collision:



Step 2 : Decide which equations to use in the problem

We know the collision is *inelastic* and there was a definite change in shape of the objects involved in the collision - there were two objects to start and after the collision there was one big mass of metal! Therefore, we know that kinetic energy is **not** conserved in the collision but total momentum **is** conserved. So:

$$\begin{array}{rcl} E_{kBefore} & \neq & E_{kAfter} \\ & \textbf{but} & \\ \overrightarrow{p_{TBefore}} & = & \overrightarrow{p_{After}} \end{array}$$

Step 3 : Solve for and quote the final velocity

So we must use conservation of momentum to solve this problem. Take the negative x-direction to have a *negative* sign:

$$\begin{array}{rcl} \overrightarrow{p_{TBefore}} & = & \overrightarrow{p_{After}} \\ \overrightarrow{p_1} + \overrightarrow{p_2} & = & \overrightarrow{p_{After}} \\ m_1\overrightarrow{u_1} + m_2\overrightarrow{u_2} & = & (m_1 + m_2)\overrightarrow{v} \\ 0 + 500(-2) & = & (500 + 500)\overrightarrow{v} \\ -1000 & = & 1000\overrightarrow{v} \\ \overrightarrow{v} & = & -1 \text{ m.s}^{-1} \end{array}$$

Therefore,

$$\overrightarrow{v} = \underline{1 \text{ m.s}^{-1} \text{ in the } \textit{negative} \text{ x} - \text{direction.}}$$

9.2 Explosions

When an object explodes, it breaks up into more than one piece and it therefore changes its shape. Explosions occur when energy is transformed from one kind *e.g. chemical potential energy* to another *e.g. heat energy* or *kinetic energy* extremely quickly. So, like in inelastic collisions, **total kinetic energy** is **not conserved** in explosions. But **total momentum** is *always* **conserved**. Thus if the momenta of some of the parts of the exploding object are measured, we can use momentum conservation to solve the problem!

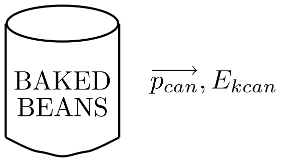
Interesting Fact:

The Tunguska event was an aerial explosion that occurred near the Podkamennaya (Stony) Tunguska River in what is now Evenkia, Siberia, at 7:17 AM on June 30, 1908. The size of the blast was later estimated to be equivalent to between 10 and 15 million tons of regular explosive. It felled an estimated 60 million trees over 2,150 square kilometers.

At around 7:15 AM, Tungus natives and Russian settlers in the hills northwest of Lake Baikal observed a huge fireball moving across the sky, nearly as bright as the Sun. A few minutes later, there was a flash that lit up half of the sky, followed by a shock wave that knocked people off their feet and broke windows up to 650 km away (*the same as the distance from Bloemfontein to Durban!*). The explosion registered on seismic stations across Europe and Asia, and produced fluctuations in atmospheric pressure strong enough to be detected in Britain. Over the next few weeks, night skies over Europe and western Russia glowed brightly enough for people to read by.

Had the object responsible for the explosion hit the Earth a few hours later, it would have exploded over Europe instead of the sparsely-populated Tunguska region, producing massive loss of human life.

In the following picture, a closed can of baked beans is put on a stove or fire:



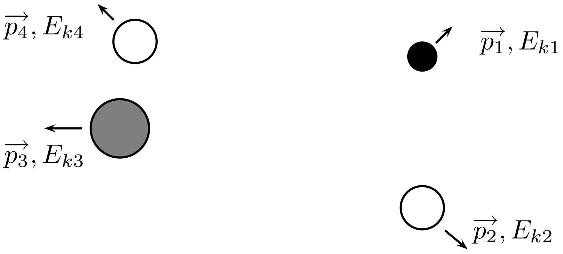
Before the can heats up and explodes, the total momentum of the system is:

$$\begin{aligned}\overrightarrow{p_{Before}} &= \overrightarrow{p_{can}} \\ &= 0\end{aligned}$$

and the total kinetic energy of the system is:

$$\begin{aligned}E_{kBefore} &= E_{kcan} \\ &= 0\end{aligned}$$

since the can isn't moving. Once the mixture of beans, juice and air inside the can reach a certain temperature, the pressure inside the can becomes so great that the can explodes! Beans and sharp pieces of metal can fly out in all directions. Energy in the system has been *transformed* from heat energy into kinetic energy.


$\overrightarrow{p_4}, E_{k4}$
$\overrightarrow{p_1}, E_{k1}$
$\overrightarrow{p_3}, E_{k3}$
$\overrightarrow{p_2}, E_{k2}$

After the explosion, the can is completely destroyed. But momentum is always conserved, so:

$$\begin{aligned}\overrightarrow{p_{Before}} &= \overrightarrow{p_{After}} \\ \overrightarrow{p_{Before}} &= \overrightarrow{p_1} + \overrightarrow{p_2} + \overrightarrow{p_3} + \overrightarrow{p_4} \\ 0 &= \overrightarrow{p_1} + \overrightarrow{p_2} + \overrightarrow{p_3} + \overrightarrow{p_4}\end{aligned}$$

However, the kinetic energy of the system is *not* conserved. The can's shape was changed in the explosion. Before the explosion the can was not moving, but after the explosion, the pieces of metal and baked beans *were* moving when they were flying out in all directions! So:

$$E_{kB} \neq E_{kA}$$

> **Safety tip:** Never heat a closed can on a stove or fire! *Always* open the can or make a hole in the lid to allow the pressure inside and outside the can to remain equal. This will prevent the can from exploding!

Worked Example 50

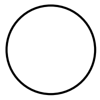
Explosions 1

Question: An object with mass $m_{Tot} = 10$ kg is sitting at rest. Suddenly it explodes into two pieces. One piece has a mass of $m_1 = 5$ kg and moves off in the negative x-direction at $\overrightarrow{v_1} = 3ms^{-1}$. What is the velocity of the other piece?

Answer:

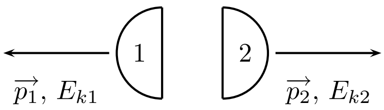
Step 1 : Draw a rough sketch of the situation

Before the explosion, the object is at rest:



$\overrightarrow{p_{Tot}} = 0$, $E_{kTot} = 0$

After the explosion, the two pieces move off:


1
2
$\overrightarrow{p_1}$, E_{k1}
$\overrightarrow{p_2}$, E_{k2}

Step 2 : Decide which equations to use in the problem

Now we know that in an explosion, total kinetic energy is *not* conserved. There is a definite change in shape of the exploding object! But we can *always* use **momentum conservation** to solve the problem. So:

$$\begin{aligned}\overrightarrow{p_{Before}} &= \overrightarrow{p_{After}} \\ \overrightarrow{p_{Before}} &= \overrightarrow{p_1} + \overrightarrow{p_2}\end{aligned}$$

But the object was initially at rest so:

$$\begin{aligned} 0 &= \overrightarrow{p_1} + \overrightarrow{p_2} \\ 0 &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2} \qquad (\mathbf{A}) \end{aligned}$$

Step 3 : Find the mass of the second piece
Now we know that m_1= 5 kg but we do not know what the mass of m_2 is. However, we do know that:

$$\begin{aligned} m_{Tot} &= m_1 + m_2 \\ m_2 &= m_{Tot} - m_1 \\ &= 10 \text{ kg} - 5 \text{ kg} \\ &= 5 \text{ kg} \end{aligned}$$

Step 4 : Solve for and quote the velocity of the other piece
Now we can substitute all the values we know into equation ($\mathbf{A}$) and solve for $\overrightarrow{v_2}$. Let's choose the positive x-direction to have a postive sign and the negative x-direction to have a negative sign:

$$\begin{aligned} 0 &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2} \qquad (\mathbf{A}) \\ 0 &= 5(-3) + 5\overrightarrow{v_2} \\ 0 &= -15 + 5\overrightarrow{v_2} \\ 5\overrightarrow{v_2} &= 15 \\ \overrightarrow{v_2} &= +3 \text{ m.s}^{-1} \end{aligned}$$

Therefore,

$$\overrightarrow{v_2} = \underline{3 \text{ m.s}^{-1} \text{ in the } \textit{positive} \text{ x} - \text{direction.}}$$

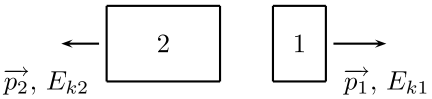
Worked Example 51

Explosions 2

Question: An object with mass m_{Tot} =15 kg is sitting at rest. Suddenly it explodes into two pieces. One piece has a mass of m_1 = 5000 g and moves off in the positive x-direction at $v_1 = 7ms^{-1}$. What is the final velocity of the other piece?
Answer:
Step 1 : Draw a rough sketch of the situation
Before the collision:

$\overrightarrow{p_{Tot}} = 0$, $E_{kTot} = 0$

After the collision:


2
1
$\overrightarrow{p_2}$, E_{k2}
$\overrightarrow{p_1}$, E_{k1}

Step 2 : Convert all units into S.I. units

$$\begin{aligned} m_1 &= 5000 \text{ g} \\ m_1 &= 5 \text{ kg} \end{aligned}$$

Step 3 : Decide which equations to use in the problem
Now we know that in an explosion, total kinetic energy is *not* conserved. There is a definite change in shape of the exploding object! But we can *always* use **momentum conservation** to solve the problem. So:

$$\begin{aligned} \overrightarrow{p_{Before}} &= \overrightarrow{p_{After}} \\ \overrightarrow{p_{Before}} &= \overrightarrow{p_1} + \overrightarrow{p_2} \end{aligned}$$

But the object was initially at rest so:

$$\begin{aligned} 0 &= \overrightarrow{p_1} + \overrightarrow{p_2} \\ 0 &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2} \qquad (\mathbf{A}) \end{aligned}$$

Step 4 : Determine the mass of the other piece
Now we know that m_1= 5 kg but we do not know what the mass of m_2 is. However, we do know that:

$$\begin{aligned} m_{Tot} &= m_1 + m_2 \\ m_2 &= m_{Tot} - m_1 \\ &= 15 \text{ kg} - 5 \text{ kg} \\ &= 10 \text{ kg} \end{aligned}$$

Step 5 : Solve for and quote the final velocity of the other piece
Now we can substitute all the values we know into equation (**A**) and solve for $\overrightarrow{v_2}$. Let's choose the positive x-direction to have a postive sign and the negative x-direction to have a negative sign:

$$\begin{aligned} 0 &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2} \qquad (\mathbf{A}) \\ 0 &= 5(7) + 10(\overrightarrow{v_2}) \\ 0 &= 35 + 10(\overrightarrow{v_2}) \\ 10(\overrightarrow{v_2}) &= -35 \\ \overrightarrow{v_2} &= -3.5 \text{ m.s}^{-1} \end{aligned}$$

Therefore,

$$\overrightarrow{v_2} = \underline{3.5 \text{ m.s}^{-1} \text{ in the } \textit{negative} \text{ x} - \text{direction.}}$$

9.3 Explosions: Energy and Heat

In explosions, you have seen that kinetic energy is not conserved. But remember that **total energy** is *always* conserved. Let's look at what happens to the energy in some more detail. If a given amount of energy is released in an explosion it is not necessarily all transformed into kinetic energy. Due to the deformation of the exploding object, often a large amount of the energy is used to break chemical bonds and heat up the pieces.

Energy is conserved but some of it is *transferred* through non-conservative processes like heating. This just means that we cannot get the energy back. It will be radiated into the environment as heat energy but it is all still accounted for.

Now we can start to mix the ideas of momentum conservation with energy transfer to make longer problems. These problems are *not* more complicated just longer. We will start off short and them combine the different ideas later on.

Long probler
should be trea
like a number
smaller proble
Focus on them
at a time.

Worked Example 52

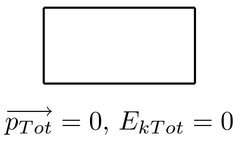
Energy Accounting 1

Question: An object with a mass of $m_t = 17$ kg explodes into two pieces of mass $m_1 = 7$ kg and $m_2 = 10$ kg. m_1 has a velocity of $9ms^{-1}$ in the negative x-direction and m_2 has a velocity of $6.3ms^{-1}$ in the positive x-direction. If the explosion released a total energy of 2000 J, how much was used in a non-conservative way?

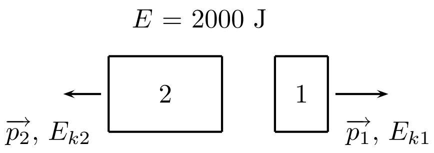
Answer:

Step 1 : Draw a rough sketch of the situation

Before the collision:



After the collision:


$E = 2000$ J
2
1
$\overrightarrow{p_2}$, E_{k2}
$\overrightarrow{p_1}$, E_{k1}

Step 2 : (NOTE TO SELF: step is deprecated, use westep instead.)

Determine what is being asked We are asked how much energy was used in a non-conservative fashion. This is the *difference* between how much energy was used in a conservative fashion and how much was used in total. We are lucky because we have everything we need to determine the kinetic energy of both pieces. The kinetic energy of the pieces is energy that was used in a conservative way.

Step 3 : Determine the total kinetic energy
The sum of the kinetic energy for the two blocks is the total kinetic energy of the pieces. So:

$$\begin{aligned}
E_{kTot} &= E_{k1} + E_{k2} \\
&= \frac{1}{2}m_1\overrightarrow{v_1}^2 + \frac{1}{2}m_2\overrightarrow{v_2}^2 \\
&= \frac{1}{2}(7)(9)^2 + \frac{1}{2}(10)(6.3)^2 \\
&= 283.5 + 198.45 \\
E_{kTot} &= 481.95 \text{ J}
\end{aligned}$$

Step 4 : Solve for and quote the final answer
The total energy that was transformed into kinetic energy is 481.95 J. We know that 2000 J of energy were released in *total.* the question makes no statements about other types of energy so we can assume that the difference was lost in a non-conservative way. Thus the total energy lost in non-conservative work is:

$$\begin{aligned}
E - E_{kTot} &= 2000 - 481.95 \\
&= 1518.05 \text{ J}
\end{aligned}$$

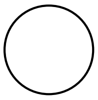
Worked Example 53

Energy Accounting 2

Question: An object at rest, with mass $m_{Tot} = 4$ kg, explodes into two pieces (m_1, m_2) with $m_1 = 2.3$ kg. m_1 has a velocity of $17ms^{-1}$ in the negative x-direction. If the explosion released a total energy of 800 J,

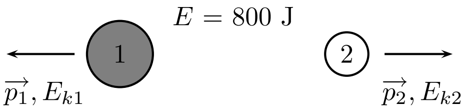
1. What is the velocity of m_2?
2. How much energy does it carry?
3. And how much energy was used in a non-conservative way?

Answer:
Step 1 : Draw a rough sketch of the situation
Before the collision:



$\overrightarrow{p_{Tot}} = 0$, $E_{kTot} = 0$

After the collision:


$E = 800$ J
1
2
$\overrightarrow{p_1}, E_{k1}$
$\overrightarrow{p_2}, E_{k2}$

Step 2 : N
ow we know that in an explosion, total kinetic energy is *not* conserved. There is a definite change in shape of the exploding object! But we can *always* use **momentum conservation** to solve the problem. So:

$$\begin{aligned}\overrightarrow{p_{Before}} &= \overrightarrow{p_{After}}\\ \overrightarrow{p_{Before}} &= \overrightarrow{p_1} + \overrightarrow{p_2}\end{aligned}$$

But the object was initially at rest so:

$$\begin{aligned}0 &= \overrightarrow{p_1} + \overrightarrow{p_2}\\ 0 &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2} \qquad (\mathbf{A})\end{aligned}$$

Step 3 : (NOTE TO SELF: step is deprecated, use westep instead.)
Now we know that m_1= 2.3 kg but we do not know what the mass of m_2 is. However, we do know that:

$$\begin{aligned}m_{Tot} &= m_1 + m_2\\ m_2 &= m_{Tot} - m_1\\ &= 4 \text{ kg} - 2.3 \text{ kg}\\ &= 1.7 \text{ kg}\end{aligned}$$

Step 4 : (NOTE TO SELF: step is deprecated, use westep instead.)
Now we can substitute all the values we know into equation ($\mathbf{A}$) and solve for $\overrightarrow{v_2}$. Let's choose the positive x-direction to have a postive sign and the negative x-direction to have a negative sign:

$$\begin{aligned}0 &= m_1\overrightarrow{v_1} + m_2\overrightarrow{v_2} \qquad (\mathbf{A})\\ 0 &= (2.3)(-17) + 1.7\overrightarrow{v_2}\\ 0 &= -39.1 + 1.7\overrightarrow{v_2}\\ 1.7\overrightarrow{v_2} &= 39.1\\ \overrightarrow{v_2} &= 23 \text{ m.s}^{-1}\end{aligned}$$

- $\underline{\overrightarrow{v_2} = 23 \text{ m.s}^{-1} \text{ in the } \textit{positive } x\text{-direction}}$

Step 5 : (NOTE TO SELF: step is deprecated, use westep instead.)
Now we need to calculate the energy that the second piece carries:

$$\begin{aligned}E_{k2} &= \frac{1}{2}m_2\overrightarrow{v_2}^2\\ &= \frac{1}{2}(1.7)(23)^2\\ &= 449.65 \text{ J}\end{aligned}$$

- The kinetic energy of the second piece is $E_{k2} = 449.65$ J

Step 6 : (NOTE TO SELF: step is deprecated, use westep instead.)
Now the amount of energy used in a non-conservative way in the explosion, is the *difference* between the amount of energy released in the explosion and the total kinetic energy of the exploded pieces:

$$E - E_{kTot} = 800 - E_{kTot}$$

We know that:

$$\begin{aligned} E_{kTot} &= E_{k1} + E_{k2} \\ &= \frac{1}{2}m_1\overrightarrow{v_1}^2 + 449.65 \\ &= \frac{1}{2}(2.3)(17)^2 + 449.65 \\ &= 332.35 + 449.65 \\ &= 782 \text{ J} \end{aligned}$$

Step 7 : (NOTE TO SELF: step is deprecated, use westep instead.)
So going back to:

$$\begin{aligned} E - E_{kTot} &= 800 - E_{kTot} \\ &= 800 - 782 \\ &= 18 \text{ J} \end{aligned}$$

- 18 J of energy was used in a non-conservative way in the explosion

9.4 Important Equations and Quantities

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
velocity	$\overrightarrow{v}$	—	$\frac{m}{s}$ **or** $m.s^{-1}$	✓
momentum	$\overrightarrow{p}$	—	$\frac{kg.m}{s}$ **or** $kg.m.s^{-1}$	✓
energy	E	J	$\frac{kg.m^2}{s^2}$ **or** $kg.m^2s^{-2}$	—

Table 9.1: Units commonly used in **Collisions and Explosions**

Momentum:

$$\overrightarrow{p} = m\overrightarrow{v} \tag{9.13}$$

Kinetic energy:

$$E_k = \frac{1}{2}m\overrightarrow{v}^2 \tag{9.14}$$

Chapter 10

Newtonian Gravitation

10.1 Properties

Gravity is a force and therefore must be described by a vector - so remember magnitude and direction. Gravity is a force that acts between *any* two objects with mass. To determine the magnitude of the force we use the following equation:

$$F = \frac{Gm_1m_2}{r^2} \tag{10.1}$$

This equation describes the force between two bodies, one of mass m_1, the other of mass m_2 (both have units of Kilogrammes, or Kg for short). The G is Newton's 'Gravitational Constant' (6.673×10^{-11} [Nm^{2}kg^{-2}]) and r is the straight line distance between the two bodies in meters.

This means the bigger the masses, the greater the force between them. Simply put, big things matter big with gravity. The $1/r^2$ factor (or you may prefer to say r^{-2}) tells us that the distance between the two bodies plays a role as well. The closer two bodies are, the stronger the gravitational force between them is. We feel the gravitational attraction of the Earth most at the surface since that is the closest we can get to it, but if we were in outer-space, we would barely even know the Earth's gravity existed!

Remember that

$$F = ma \tag{10.2}$$

which means that every object on the earth feels the same gravitational acceleration! That means whether you drop a pen or a book (from the same height), they will both take the same length of time to hit the ground... in fact they will be head to head for the entire fall if you drop them at the same time. We can show this easily by using the two equations above (10.1 and 10.2). The force between the Earth (which has the mass m_e) and an object of mass m_o is

$$F = \frac{Gm_om_e}{r^2} \tag{10.3}$$

and the acceleration of an object of mass m_o (in terms of the force acting on it) is

$$a_o = \frac{F}{m_o} \tag{10.4}$$

So we substitute equation (10.3) into equation (10.4), and we find that

$$a_o = \frac{Gm_e}{r^2} \tag{10.5}$$

Since it doesn't matter what m_o is, this tells us that the acceleration on a body (due to the Earth's gravity) does not depend on the mass of the body. Thus all objects feel the same gravitational acceleration. The force on different bodies will be different but the acceleration will be the same. Due to the fact that this acceleration caused by gravity is the same on all objects we label it differently, instead of using a we use g which we call the gravitational acceleration.

10.2 Mass and Weight

Weight is a force which is measured in Newtons, it is the force of gravity on an object. People are always asking other people "What is your weight?" when in fact they should be asking "What is your mass?".

Mass is measured in Kilograms (Kg) and is the amount of matter in an object, it doesn't change unless you add or remove matter from the object (if you continue to study physics through to university level, you will find that Einstein's theory of relativity means that mass can change when you travel as fast as light does, but you don't need to worry about that right now). There are 1000g in 1Kg and 1000Kg in a Tonne.

To change mass into weight we use Newton's 2nd Law which is F = Ma. The weight is the force and gravity the acceleration, it can be rewritten as:

$$W = mg \tag{10.6}$$

W is the Weight, measured in Newtons. M is the Mass, measured in Kg and g is the acceleration due to gravity, measured in m/s^2 it is equal to 10 on the Earth.

10.2.1 Examples

1. A bag of sugar has a mass of 1Kg, what is it's weight? (Acceleration due to gravity = $10m/s^2$)

    - Step 1: Always write out the equation, it helps you to understand the question, and you will get marks as well.
    $$W = Mg \tag{10.7}$$
    - Step 2: Fill in all the values you know. (remember to make sure the mass is in Kg and NOT in grams or Tonnes!)
    $$\begin{aligned} W &= 1 \times 10 \\ W &= 10 \end{aligned} \tag{10.8}$$
    - Step 3: Write out the answer remembering to include the units! You will lose marks if you don't
    $$W = 10\text{Newtons} \tag{10.9}$$

2. A space-man has a mass of 90Kg, what is his weight (a) on the earth? (b) on the moon? (c) in outer space? (The acceleration due to gravity on the earth is $10m/s^2$, on the moon gravity is 1/6 of the gravity on earth).

    (a)
    $$\begin{aligned} W &= Mg \\ W &= 90 \times 10 = 900 \\ W &= 900\text{Newtons} \end{aligned} \tag{10.10}$$

(b)

$$\begin{aligned} W &= Mg \\ W &= 90 \times 10 \times 1/6 = 150 \\ W &= 150\text{Newtons} \end{aligned} \tag{10.11}$$

(c) Weightless in outer space because $g = 0$.

So now when somebody asks you your weight, you know to reply “Anything!! But my mass is a different matter!”

10.3 Normal Forces

If you put a book on a table it does not accelerate it just lies on the table. We know that gravity is acting on it with a force

$$F = G\frac{m_E m_{book}}{r^2} \tag{10.12}$$

but if there is a net force there MUST be an acceleration and there isn’t. This means that the gravitational force is being balanced by another force[1].

This force we call the normal force. It is the reaction force between the book and the table. It is equal to the force of gravity on the book. This is also the force we measure when we measure the weight of something.

The most interesting and illustrative normal force question, that is often asked, has to do with a scale in a lift. Using Newton’s third law we can solve these problems quite easily.

When you stand on a scale to measure your weight you are pulled down by gravity. There is no acceleration downwards because there is a reaction force we call the normal force acting upwards on you. This is the force that the scale would measure. If the gravitational force were less then the reading on the scale would be less.

Worked Example 54

Normal Forces 1

Question: A man weighing $100kg$ stands on a scale (measuring newtons). What is the reading on the scale?
Answer:
Step 1 : Decide what information is supplied
We are given the mass of the man. We know the gravitational acceleration that acts on him - $g = 10m/s^2$.
Step 2 : Decide what equation to use to solve the problem
The scale measures the normal force on the man. This is the force that balances gravity. We can use Newton’s laws to solve the problem:

$$F_r = F_g + F_N \tag{10.13}$$

where F_r is the resultant force on the man.

[1] Newton’s third law!

Step 3 : Firstly we determine the net force acting downwards on the man due to gravity

$$\begin{aligned} F_g &= mg \\ &= 100kg \times 9.8\frac{m}{s^2} \\ &= 980\frac{kgm}{s^2} \\ &= 980N \ downwards \end{aligned}$$

Step 4 : Now determine the normal force acting upwards on the man
We now know the gravitational force downwards. We know that the sum of all the forces must equal the resultant acceleration times the mass. The overall resultant acceleration of the man on the scale is 0 - so $F_r = 0$.

$$\begin{aligned} F_r &= F_g + F_N \\ 0 &= -980N + F_N \end{aligned}$$

$$F_N = 980N \ upwards$$

Step 5 : Quote the final answer
The normal force is then $980N$ upwards. It exactly balances the gravitational force downwards so there is no net force and no acceleration on the man.

Now we are going to add things to exactly the same problem to show how things change slightly. We will now move to a lift moving at constant velocity. Remember if velocity is constant then acceleration is zero.

Worked Example 55

Normal Forces 2

Question: A man weighing $100kg$ stands on a scale (measuring newtons) inside a lift moving downwards at $2\frac{m}{s}$. What is the reading on the scale?
Answer:
Step 1 : Decide what information is supplied
We are given the mass of the man and the acceleration of the lift. We know the gravitational acceleration that acts on him.
Step 2 : Decide which equation to use to solve the problem
Once again we can use Newton's laws. We know that the sum of all the forces must equal the resultant acceleration times the mass (This is the resultant force, F_r).

$$F_r = F_g + F_N \tag{10.14}$$

Step 3 : Firstly we determine the net force acting downwards on the man due to gravity

$$\begin{aligned} F_g &= mg \\ &= 100kg \times 9.8\frac{m}{s^2} \\ &= 980\frac{kgm}{s^2} \\ &= 980N\ downwards \end{aligned}$$

Step 4 : Now determine the normal force acting upwards on the man
The scale measures this normal force, so once we've determined it we will know the reading on the scale. Because the lift is moving at constant velocity the overall resultant acceleration of the man on the scale is 0. If we write out the equation:

$$\begin{aligned} F_r &= F_g + F_N \\ 0 &= -980N + F_N \end{aligned}$$

$$F_N = 980N\ upwards$$

Step 5 : Quote the final answer
The normal force is then $980N$ upwards. It exactly balances the gravitational force downwards so there is no net force and no acceleration on the man.

In this second example we get exactly the same result because the net acceleration on the man was zero! If the lift is accelerating downwards things are slightly different and now we will get a more interesting answer!

Worked Example 56

Normal Forces 3

Question: A man weighing $100kg$ stands on a scale (measuring newtons) inside a lift accelerating downwards at $2\frac{m}{s^2}$. What is the reading on the scale?
Answer:
Step 1 : Decide what information is supplied
We are given the mass of the man and his resultant acceleration - this is just the acceleration of the lift. We know the gravitational acceleration that acts on him.
Step 2 : Decide which equation to use to solve the problem
Once again we can use Newton's laws. We know that the sum of all the forces must equal the resultant acceleration times the mass (This is the resultant force, F_r).

$$F_r = F_g + F_N \tag{10.15}$$

Step 3 : Firstly we determine the net force acting downwards on the man due to gravity, F_g

$$\begin{aligned} F_g &= mg \\ &= 100kg \times 9.8\frac{m}{s^2} \\ &= 980\frac{kgm}{s^2} \\ &= 980N\ downwards \end{aligned}$$

Step 4 : Now determine the normal force acting upwards on the man, F_N

We know that the sum of all the forces must equal the resultant acceleration times the mass. The overall resultant acceleration of the man on the scale is $2\frac{m}{s^2}$ downwards. If we write out the equation:

$$\begin{aligned} F_r &= F_g + F_N \\ 100kg \times (-2)\frac{m}{s^2} &= -980N + F_N \\ -200\frac{kgm}{s^2} &= -980N + F_N \\ -200N &= -980N + F_N \\ F_N = 780N\ upwards & \end{aligned}$$

Step 5 : Quote the final answer

The normal force is then $780N$ upwards. It balances the gravitational force downwards just enough so that the man only accelerates downwards at $2\frac{m}{s^2}$.

Worked Example 57

Normal Forces 4

Question: A man weighing $100kg$ stands on a scale (measuring newtons) inside a lift accelerating upwards at $4\frac{m}{s^2}$. What is the reading on the scale?

Answer:

Step 1 : Decide what information is supplied

We are given the mass of the man and his resultant acceleration - this is just the acceleration of the lift. We know the gravitational acceleration that acts on him.

Step 2 : Decide which equation to use to solve the problem

Once again we can use Newton's laws. We know that the sum of all the forces must equal the resultant acceleration times the mass (This is the resultant force, F_r).

$$F_r = F_g + F_N \tag{10.16}$$

Step 3 : Firstly we determine the net force acting downwards on the man due to gravity, F_g

$$\begin{aligned} F_g &= mg \\ &= 100kg \times 9.8\frac{m}{s^2} \\ &= 980\frac{kgm}{s^2} \\ &= 980N\ downwards \end{aligned}$$

Step 4 : Now determine the normal force upwards, F_N

We know that the sum of all the forces must equal the resultant acceleration times the mass. The overall resultant acceleration of the man on the scale is $2\frac{m}{s^2}$ downwards. if we write out the equation:

$$\begin{aligned} F_r &= F_g + F_N \\ 100kg \times (4)\frac{m}{s^2} &= -980N + F_N \\ 400\frac{kgm}{s^2} &= -980N + F_N \\ 400N &= -980N + F_N \\ F_N = 1380N\ upwards & \end{aligned}$$

Step 5 : Quote the final answer

The normal force is then $1380N$ upwards. It balances the gravitational force and then in addition applies sufficient force to accelerate the man upwards at $4\frac{m}{s^2}$.

10.4 Comparative problems

Here always work with multiplicative factors to find something new in terms of something old.

Worked Example 58

Comparative Problem 1

Question: On Earth a man weighs $70kg$. Now if the same man was instantaneously beamed to the planet Zirgon, which has the same size as the Earth but twice the mass, what would he weigh? (NOTE TO SELF: Vanessa: isn't this confusing weight and mass?)

Answer:

Step 1 : We start with the situation on Earth

$$W = mg = G\frac{m_E m}{r^2} \tag{10.17}$$

Step 2 : Now we consider the situation on Zirgon

$$W_Z = mg_Z = G\frac{m_Z m}{r_Z^2} \tag{10.18}$$

Step 3 : Relation between conditions on Earth and Zirgon
but we know that $m_Z = 2m_E$ and we know that $r_Z = r$ so we could write the equation again and substitute these relationships in:
Step 4 : Substitute

$$W_Z = mg_Z = G\frac{(2m_E)m}{(r)^2} \tag{10.19}$$

$$W_Z = 2(G\frac{(m_E)m}{(r)^2}) \tag{10.20}$$

Step 5 : Relation between weight on Zirgon and Earth

$$W_Z = 2(W) \tag{10.21}$$

Step 6 : Quote the final answer
so on Zirgon he weighs 140kg.

10.4.1 Principles

- Write out first case
- Write out all relationships between variable from first and second case
- Write out second case
- Substitute all first case variables into second case
- Write second case in terms of first case

Interesting Fact:

The acceleration due to gravity at the Earth's surface is, by convention, equal to 9.80665 ms^{-2}. (The actual value varies slightly over the surface of the Earth). This quantity is known as g. The following is a list of the gravitational accelerations (in multiples of g) at the surfaces of each of the planets in our solar system:

Mercury	0.376
Venus	0.903
Earth	1
Mars	0.38
Jupiter	2.34
Saturn	1.16
Uranus	1.15
Neptune	1.19
Pluto	0.066

Note:
The "surface" is taken to mean the cloud tops of the gas giants (Jupiter, Saturn, Uranus and Neptune) in the above table.

Worked Example 59

Comparative Problem 2

Question: On Earth a man weighs $70kg$. On the planet Beeble how much will he weigh if Beeble has mass half of that of the Earth and a radius one quarter that of the earth.

Answer:

Step 1 : Start with the situation on Earth

$$W = mg = G\frac{m_E m}{r^2} \tag{10.22}$$

Step 2 : Now consider the situation on Beeble

$$W_B = mg_B = G\frac{m_B m}{r_B^2} \tag{10.23}$$

Step 3 : Relation between conditions on Earth and on Beeble

We know that $m_B = \frac{1}{2}m_E$ and we know that $r_B = \frac{1}{4}r$ so we could write the equation again and substitute these relationships in:

Step 4 : Substitute

$$\begin{aligned} W_B &= mg_B = G\frac{(m_B)m}{(r_B)^2} \\ &= mg_B = G\frac{(\frac{1}{2}m_E)m}{(\frac{1}{4}r)^2} \\ &= 8(G\frac{(m_E)m}{(r)^2}) \end{aligned}$$

Step 5 : Relation between weight on Earth and weight on Beeble

$$W_B = 8(W) \tag{10.24}$$

Step 6 : Quote the final answer

So the man weighs $560kg$ on Beeble!

Interesting Fact: **Did you know** that the largest telescope in the Southern Hemisphere is the South African Large Telescope (SALT) which came online in 2005 outside Sutherland in the Karoo.

10.5 Falling bodies

Objects on the earth fall because there is a gravitation force between them and the earth - which results in an acceleration - as we saw above. So if you hold something in front of you and let it go - it will fall.

It falls because of an acceleration toward the centre of the earth which results from the gravitational force between the two.

These bodies move in a straight line from the point where they start to the centre for the earth. This means we can reuse everything we learnt in rectilinear motion. the only thing that needs thinking about is the directions we are talking about.

We need to choose either up or down as positive just like we had to choose a positive direction in standard rectilinear motion problems. this is the hardest part. If you can do rectilinear motion you can do falling body problems. Just remember the acceleration they feel is constant and because of gravity - but once you have chosen your directions you can forget that gravity has anything to do with the problem - all you have is a rectilinear motion problem with a constant acceleration!!

10.6 Terminal velocity

Physics is all about being simple - all we do is look at the world around us and notice how it really works. It is the one thing everyone is qualified to do - we spend most of our time when we are really young experimenting to find out how things work.

Take a book - wave it in the air - change the angle and direction. what happens of course there is resistance. different angles make it greater - the faster the book moves the greater it is. The bigger the area of the book moving in the direction of motion the greater the force.

So we know that air resistance exists! it is a force. So what happens when an object falls? of course there is air resistance - or drag as it is normally called. There is an approximate formula for the drag force as well.

The important thing to realise is that when the drag force and the gravitational force are equal for a falling body there is no net force acting on it - which means no net acceleration. That does not mean it does not move - but it means that its speed does not change.

It falls at a constant velocity! This velocity is called terminal velocity.

10.7 Drag force

The actual force of air resistance is quite complicated. Experiment by moving a book through the air with the face of the book and then the side of the book forward, you will agree that the area of the book makes a difference as to how much you must work in order to move the book at the same speed in both cases. This is why racing cars are slim-lined in design, and not shaped like a big box!

Get a plastic container lid (or anything waterproof) swing it around in air and then try to swing it around under water. The density of the water is much larger than the air, making you have to work harder at swinging the lid in water. This is why boats and submarines are a lot slower than aeroplanes!

So we know that density, area and speed all play a role in the drag force. The expression we use for drag force is

$$D = \frac{1}{2}C\rho Av^2 \tag{10.25}$$

where C is a constant which depends on the object and fluid interactions, ρ is the density, A is the area and v is the velocity.

10.8 Important Equations and Quantities

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
mass	m	—	kg **or** —	—
velocity	$\overrightarrow{v}$	—	$\frac{m}{s}$ **or** $m.s^{-1}$	✓
force	$\overrightarrow{F}$	N	$\frac{kg.m}{s^2}$ **or** $kg.m.s^{-2}$	✓
energy	E	J	$\frac{kg.m^2}{s^2}$ **or** $kg.m^2.s^{-2}$	—

Table 10.1: Units used in **Newtonian Gravitation**

Chapter 11

Pressure

11.1 Important Equations and Quantities

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
			or	

Table 11.1: Units used in **Pressure**

Essay 3: Pressure and Forces

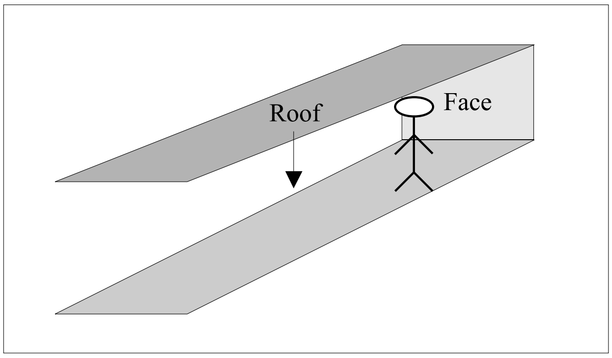
Author: Asogan Moodaly

Asogan Moodaly received his Bachelor of Science degree (with honours) in Mechanical Engineering from the University of Natal, Durban in South Africa. For his final year design project he worked on a 3-axis filament winding machine for composite (Glass re-enforced plastic in this case) piping. He worked in Vereeniging, Gauteng at Mine Support Products (a subsidiary of Dorbyl Heavy Engineering) as the design engineer once he graduated. He currently lives in the Vaal Triangle area and is working for Sasol Technology Engineering as a mechanical engineer, ensuring the safety and integrity of equipment installed during projects.

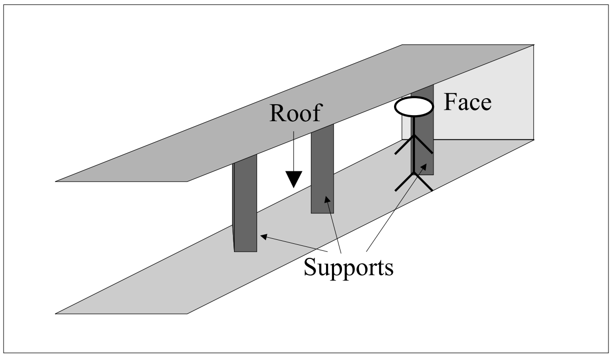
Pressure and Forces

In the mining industry, the roof (hangingwall) tends to drop as the face of the tunnel (stope) is excavated for rock containing gold.

As one can imagine, a roof falling on one's head is not a nice prospect! Therefore the roof needs to be supported.

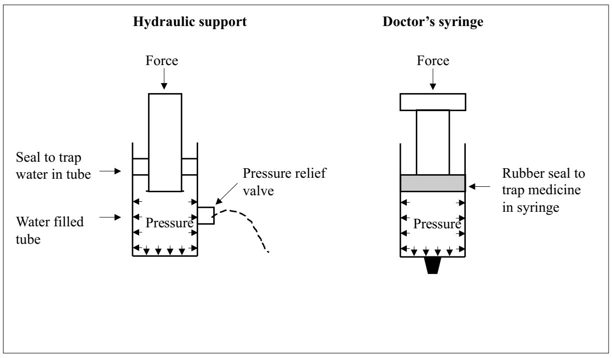


The roof is not one big uniform chunk of rock. Rather it is broken up into smaller chunks. It is assumed that the biggest chunk of rock in the roof has a mass of less than 20 000 kgs therefore each support has to be designed to resist a force related to that mass. The strength of the material (either wood or steel) making up the support is taken into account when working out the minimum required size and thickness of the parts to withstand the force of the roof.

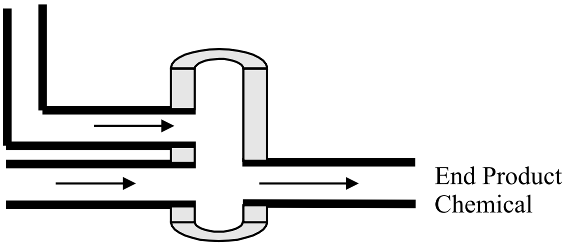


Sometimes the design of the support is such that the support needs to withstand the rock mass without the force breaking the roof..

Therefore hydraulic supports (hydro = water) use the principles of force and pressure such that as a force is exerted on the support, the water pressure increases. A pressure relief valve then squirts out water when the pressure (and thus the force) gets too large. Imagine a very large, modified doctor's syringe.



In the petrochemical industry, there are many vessels and pipes that are under high pressures. A vessel is a containment unit (Imagine a pot without handles, that has the lid welded to the pot that would be a small vessel) where chemicals mix and react to form other chemicals, amongst other uses.



The end product chemicals are sold to companies that use these chemicals to make shampoo, dishwashing liquid, plastic containers, fertilizer, etc. Anyway, some of these chemical reactions require high temperatures and pressures in order to work. These pressures result in forces being applied to the insides of the vessels and pipes. Therefore the minimum thickness of the pipe and vessels walls must be determined using calculations, to withstand these forces. These calculations take into account the strength of the material (typically steel, plastic or composite), the diameter and of course the pressure inside the equipment. Let examine the concepts of force and pressure in further detail.

Chapter 12

Heat and Properties of Matter

12.1 Phases of matter

12.1.1 Density

Matter is a substance which has mass and occupies space. The density of matter refers to how much mass is in a given volume. Said differently, you can imagine the density to be the amount of mass packed into a given volume.

$$density = \frac{Mass}{Volume}$$

If we consider a bar of soap and a bar of steel with the same volume, the steel will have more mass because it has a greater density. The density is greater in steal because more atoms are closely packed in comparison to the soap. Although they are both the same size, the bar of steel will be "heavier" because it has more mass.

Worked Example 60

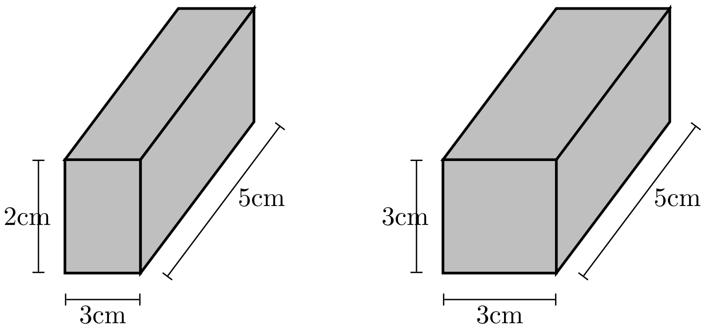
Density of objects

A bar of aluminum (Al) has dimensions 2cm x 3cm x 5cm with a mass of 81g. A bar of lead (Pb) has dimensions 3cm x 3cm x 5cm and a mass of 510.3g. Calculate the density of the aluminum and lead.
Solution:
First we calculate the volume of Al and Pb:

$$volume = Length * Width * Height$$

For Aluminum: $volume = 2cm * 3cm * 5cm = 30cm^3$
For Lead: $volume = 3cm * 3cm * 5cm = 45cm^3$

We can now calculate the densities using the mass and volume of each material.
For Aluminum: $density = \frac{81g}{30cm^3} = 2.7g/cm^3$
For Lead: $density = \frac{510.3g}{45cm^3} = 11.34g/cm^3$



Now that you know the density of aluminum and lead, which object would be bigger (larger volume): 1kg of Lead or 1kg of Aluminum.

Solution:
1kg of aluminum will be much larger in volume than 1kg of lead. Aluminum has a smaller density so it will take a lot more of it to have a weight of 1kg. Lead is much more dense, so it will take less for it to weigh 1kg.

The density of liquids and gases can be calculated the same way as in solids. If the mass and volume of a liquid is known, the density can be calculated. We can often determine which liquid has a greater density by mixing two liquids and seeing how they settle. The more dense liquid will fall towards the bottom, or 'sink'. If you have ever added olive oil to water, you have seen it sits on the surface, or 'floats'. This is because olive oil is less dense than water. Fog occurs when water vapor becomes more dense than air("a cloud that sinks in air").

This principle can be used with solids and liquids. In fact, it is the density of an object that determines if it will float or sink in water. Objects with densities greater than water will sink.

Worked Example 61

Objects floating in water

Ivory soap is famous for "soap that floats". If a 5cm x 3cm x 10cm bar of ivory soap weighs 1.35 Newtons, show that its density is less than water.

Solution:
First calculate the bars volume: $volume = 3cm * 5cm * 10cm = 150cm^3$
Now we must determine the mass of the bar based on its weight. We will use Newton's Second law ($F = ma$):

$$Weight = mass * gravity \quad \Longrightarrow \quad Weight = 9.8m/s^2 * Mass$$

$$Mass = \frac{1.35N}{9.8m/s^2} = .138kg$$

Using the mass and the volume we determine the density of the soap:

$$density = \frac{138g}{150cm^3} = .92g/cm^3$$

Water has a density of $1g/cm^3$, therefore the soap is less dense than water, allowing it to float.

12.2 Phases of matter

Although phases are conceptually simple, they are hard to define precisely. A good definition of a phase of a system is a region in the parameter space of the system's thermodynamic variables in which the free energy is analytic. Equivalently, two states of a system are in the same phase if they can be transformed into each other without abrupt changes in any of their thermodynamic properties.

All the thermodynamic properties of a system – the entropy, heat capacity, magnetization, compressibility, and so forth – may be expressed in terms of the free energy and its derivatives. For example, the entropy is simply the first derivative of the free energy with temperature. As long as the free energy remains analytic, all the thermodynamic properties will be well-behaved.

When a system goes from one phase to another, there will generally be a stage where the free energy is non-analytic. This is known as a phase transition. Familiar examples of phase transitions are melting (solid to liquid), freezing (liquid to solid), boiling (liquid to gas), and condensation (gas to liquid). Due to this non-analyticity, the free energies on either side of the transition are two different functions, so one or more thermodynamic properties will behave very differently after the transition. The property most commonly examined in this context is the heat capacity. During a transition, the heat capacity may become infinite, jump abruptly to a different value, or exhibit a "kink" or discontinuity in its derivative.

In practice, each type of phase is distinguished by a handful of relevant thermodynamic properties. For example, the distinguishing feature of a solid is its rigidity; unlike a liquid or a gas, a solid does not easily change its shape. Liquids are distinct from gases because they have much lower compressibility: a gas in a large container fills the container, whereas a liquid forms a puddle in the bottom. Many of the properties of solids, liquids, and gases are not distinct; for instance, it is not useful to compare their magnetic properties. On the other hand, the ferromagnetic phase of a magnetic material is distinguished from the paramagnetic phase by the presence of bulk magnetization without an applied magnetic field.

To take another example, many substances can exist in a variety of solid phases each corresponding to a unique crystal structure. These varying crystal phases of the same substance are called polymorphs. Diamond and graphite are examples of polymorphs of carbon. Graphite is composed of layers of hexagonally arranged carbon atoms, in which each carbon atom is strongly bound to three neighboring atoms in the same layer and is weakly bound to atoms in the neighboring layers. By contrast in diamond each carbon atom is strongly bound to four neighboring carbon atoms in a cubic array. The unique crystal structures of graphite and diamond are responsible for the vastly different properties of these two materials.

Metastable phases

Metastable states may sometimes be considered as phases, although strictly speaking they aren't because they are unstable. For example, each polymorph of a given substance is usually only stable over a specific range of conditions. For example, diamond is only stable at extremely high pressures. Graphite is the stable form of carbon at normal atmospheric pressures. Although diamond is not stable at atmospheric pressures and should transform to graphite, we know that diamonds exist at these pressures. This is because at normal temperatures the transformation from diamond to graphite is extremely slow. If we were to heat the diamond, the rate of transformation would increase and the diamond would become graphite. However, at normal temperatures the diamond can persist for a very long time.

Another important example of metastable polymorphs occurs in the processing of steel. Steels are often subjected to a variety of thermal treatments designed to produce various combinations of stable and metastable iron phases. In this way the steel properties, such as hardness and strength can be adjusted by controlling the relative amounts and crystal sizes of the various phases that form.

Phase diagrams

The different phases of a system may be represented using a phase diagram. The axes of the diagrams are the relevant thermodynamic variables. For simple mechanical systems, we generally use the pressure and temperature. The following figure shows a phase diagram for a typical material exhibiting solid, liquid and gaseous phases.

The markings on the phase diagram show the points where the free energy is non-analytic. The open spaces, where the free energy is analytic, correspond to the phases. The phases are separated by lines of non-analyticity, where phase transitions occur, which are called phase boundaries.

In the above diagram, the phase boundary between liquid and gas does not continue indefinitely. Instead, it terminates at a point on the phase diagram called the critical point. This reflects the fact that, at extremely high temperatures and pressures, the liquid and gaseous phases become indistinguishable. In water, the critical point occurs at around 647 K (374 C or 705 F) and 22.064 MPa.

The existence of the liquid-gas critical point reveals a slight ambiguity in our above definitions. When going from the liquid to the gaseous phase, one usually crosses the phase boundary, but it is possible to choose a path that never crosses the boundary by going to the right of the critical point. Thus, phases can sometimes blend continuously into each other. We should note, however, that this does not always happen. For example, it is impossible for the solid-liquid phase boundary to end in a critical point in the same way as the liquid-gas boundary, because the solid and liquid phases have different symmetry.

An interesting thing to note is that the solid-liquid phase boundary in the phase diagram of most substances, such as the one shown above, has a positive slope. This is due to the solid phase having a higher density than the liquid, so that increasing the pressure increases the melting temperature. However, in the phase diagram for water the solid-liquid phase boundary has a negative slope. This reflects the fact that ice has a lower density than water, which is an unusual property for a material.

12.2.1 Solids, liquids, gasses

12.2.2 Pressure in fluids

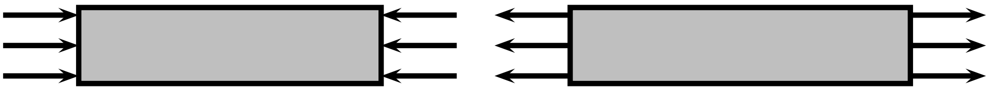
12.2.3 change of phase

12.3 Deformation of solids

12.3.1 strain, stress

Stress (σ) and strain (ϵ) is one of the most fundamental concepts used in the mechanics of materials. The concept can be easily illustrated by considering a solid, straight bar with a constant cross section throughout its length where a force is distributed evenly at the ends of the bar. This force puts a stress upon the bar. Like pressure, the stress is the force per unit area. In this case the area is the cross sectional area of the bar.

$$stress = \frac{Force}{Area_{crosssection}} \quad \Longrightarrow \quad \sigma = \frac{F}{A}$$



(A) Bar under compression

(B) Bar under tension

Figure 12.1: Illustration of Bar

The bar in figure 1a is said to be under compression. If the direction of the force ($\overrightarrow{F}$), were reversed, stretching the bar, it would be under tension (fig. 1b). Using intuition, you can imagine how the bar might change in shape under compression and tension. Under a compressive load, the bar will shorten and thicken. In contrast, a tensile load will lengthen the bar and make it thinner.

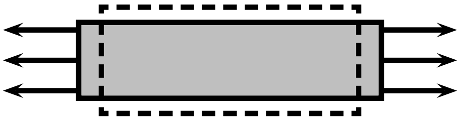


Figure 12.2: Bar changes length under tensile stress

For a bar with an original length L, the addition of a stress will result in change of length $\triangle L$. With $\triangle L$ and L we can now define strain as the ratio between the two. That is, strain is defined as the fractional change in length of the bar:

$$Strain \quad \equiv \quad \frac{\triangle L}{L}$$

12.3.2 Elastic and plastic behavior

Material properties are often characterized by a stress versus strain graph (figure x.xx). One way in which these graphs can be determined is by tensile testing. In this process, a machine

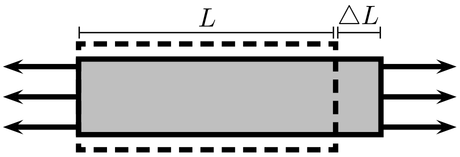


Figure 12.3: Left end of bar is fixed as length changes

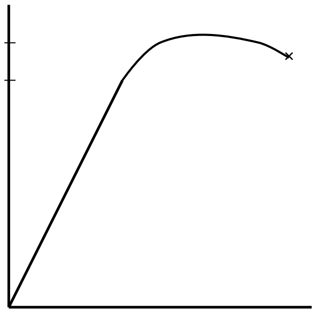


Figure 12.4: dashed line represents plastic recovery **incomplete**

stretches a the material by constant amounts and the corresponding stress is measured and plotted. Typical solid metal bars will show a result like that of figure x.xx. This is called a Type II response. Other materials may exibit different responses. We will only concern ourself with Type II materials.

The linear region of the graph is called the elastic region. By obtaining the slope of the linear region, it is easy to find the strain for a given stress, or vice-versa. This slope shows itself to be very useful in characterizing materials, so it is called the Modulus of Elasticity, or Young's Modulus:

$$E = \frac{stress}{strain} = \frac{F/A}{\Delta L/L}$$

The elastic region has the unique property that allows the material to return to its original shape when the stress is removed. As the stress is removed it will follow line back to zero. One may think of stretching a spring and then letting it return to its original length. When a stress is applied in the linear region, the material is said to undergo elastic deformation.

When a stress is applied that is in the non-linear region, the material will no longer return to its original shape. This is referred to as plastic deformation. If you have overstretched a spring you have seen that it no longer returns to its initial length; it has been plastically deformed. The stress where plastic behavior begins is called the yield strength (point A, fig x).

When a material has plastically deformed it will still recover some of its shape (like an overstretched spring). When a stress in the non-linear region is removed, the stress strain graph will follow a line with a slope equal to the modulus of elasticity (see the dashed line in figure x.xx). The plastically deformed material will now have a linear region that follows the dashed line.

Greater stresses in the plastic region will eventually lead to fracture (the material breaks). The maximum stress the material can undergo before fracture is the ultimate strength.

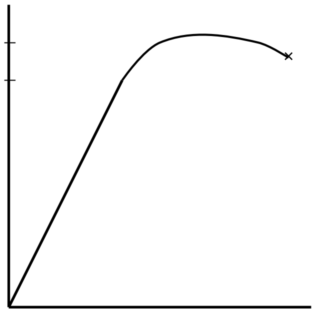


Figure 12.5: dashed line represents plastic recovery **incomplete**

12.4 Ideal gasses

Author: Gérald Wigger

Gérald Wigger started his Physics studies at ETH in Zuerich, Switzerland. He moved to Cape Town, South Africa, for his Bachelor of Science degree (with honours) in Physics from the University of Cape Town in 1998. Returned to Switzerland, he finished his Diploma at ETH in 2000 and followed up with a PhD in the Solid State Physics group of Prof. Hans-Ruedi Ott at ETH. He graduated in the year 2004. Being awarded a Swiss fellowship, he moved to Stanford University where he is currently continuing his Physics research in the field of Materials with novel electronic properties.

Any liquid or solid material, heated up above its boiling point, undergoes a transition into a gaseous state. For some materials such as aluminium, one has to heat up to three thousand degrees Celsius (°C), whereas Helium is a gas already at -269 °C. For more examples see Table 12.1. As we find very strong bonding between the atoms in a solid material, a gas consists of molecules which do interact very poorly. If one forgets about any electrostatic or intermolecular attractive forces between the molecules, one can assume that all collisions are perfectly elastic. One can visualize the gas as a collection of perfectly hard spheres which collide but which otherwise do not interact with each other. In such a gas, all the internal energy is in the form of kinetic energy and any change in internal energy is accompanied by a change in temperature. Such a gas is called an **ideal gas**.

In order for a gas to be described as an ideal gas, the temperature should be raised far enough above the melting point. A few examples of ideal gases at room temperature are Helium, Argon and hydrogen. Despite the fact that there are only a few gases which can be accurately described as an ideal gas, the underlying theory is widely used in Physics because of its beauty and simplicity.

A thermodynamic system may have a certain substance or material whose quantity can be expressed in mass or mols in an overall volume. These are *extensive* properties of the system. In the following we will be considering often *intensive* versus extensive quantities. A material's intensive property, is a quantity which does not depend on the size of the material, such as temperature, pressure or density. Extensive properties like volume, mass or number of atoms on

Material	Temperature in Celsius	Temperature in Kelvin
Aluminium	2467 °C	2740 K
Water	100 °C	373.15 K
Ethyl alcohol	78.5 °C	351.6 K
Methyl ether	-25 °C	248 K
Nitrogen	-195.8 °C	77.3 K
Helium	-268.9 °C	4.2 K

Table 12.1: Boiling points for various materials in degrees Celsius and in Kelvin

quantity	unit	intensive or extensive
pressure p	Pa	intensive
volume V	m^3	extensive
molar volume v_{mol}	$\mathrm{m}^3/\mathrm{mol}$	intensive
temperature T	K	intensive
mass M	kg	extensive
density ρ	kg/m^3	intensive
internal energy E	J	extensive

Table 12.2: Intensive versus extensive properties of matter

the other hand gets bigger the bigger the material is (see Table 12.2 for various intensive/extensive properties). If the substance is evenly distributed throughout the volume in question, then a value of volume per amount of substance may be used as an intensive property. For an example, for an amount called a mol, volume per mol is typically called molar volume. Also, a volume per mass for a specific substance may be called specific volume. In the case of an ideas gas, a simple equation of state relates the three intensive properties, temperature, pressure, and molar or specific volume. Hence, for a closed system containing an ideal gas, the state can be specified by giving the values of any two of pressure, temperature, and molar volume.

12.4.1 Equation of state

The ideal gas can be described with a single equation. However, in order to arrive there, we will be introducing three different equations of state, which lead to the ideal gas law. The combination of these three laws leads to a complete picture of the ideal gas.

1661 - Robert Boyle used a U-tube and Mercury to develop a mathematical relationship between pressure and volume. To a good approximation, the pressure and volume of a fixed amount of gas at a constant temperature were related by

$$p \cdot V = constant$$

$$\begin{array}{ll} p & : \text{pressure } (Pa) \\ V & : \text{Volume } (m^3) \end{array}$$

In other words, if we compress a given quantity of gas, the pressure will increase. And if we put it under pressure, the volume of the gas will decrease proportionally.

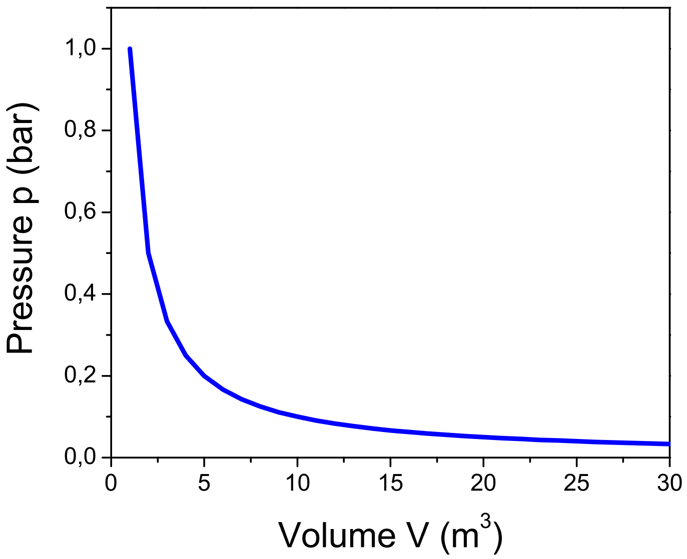


Figure 12.6: Pressure-Volume diagram for the ideal gas at constant temperature.

Worked Example 62

compressed Helium gas

A sample of Helium gas at 25°C is compressed from 200 cm^3 to 0.240 cm^3. Its pressure is now 3.00 cm Hg. What was the original pressure of the Helium?
Solution:
It's always a good idea to write down the values of all known variables, indicating whether the values are for initial or final states. Boyle's Law problems are essentially special cases of the Ideal Gas Law:
Initial: p_1 = ?; V_1 = 200 cm^3;
Final: p_2 = 3.00 cm Hg; V_2 = 0.240 cm^3;
Since the number of molecules stays constant and the temperature is not changed along the process, so

$$p_1 \cdot V_1 = p_2 \cdot V_2$$

hence

$$p_1 = p_2 \cdot V_2/V_1 = 3.00cmHg \cdot 0.240cm^3/200cm^3$$

Setting in the values yields $p_1 = 3.60 \cdot 10^{-3}$ cm Hg.
Did you notice that the units for the pressure are in cm Hg? You may wish to convert this to a more common unit, such as millimeters of mercury, atmospheres, or pascals.
$3.60 \cdot 10^{-3}$ Hg · 10mm/1 cm = $3.60 \cdot 10^{-2}$ mm Hg

$3.60 \cdot 10^{-3}$ Hg $\cdot$ 1 atm/76.0 cm Hg $= 4.74 \cdot 10^{-5}$ atm

One way to experience this is to dive under water. There is air in your middle ear, which is normally at one atmosphere of pressure to balance the air outside your ear drum. The water will put pressure on the ear drum, thereby compressing the air in your middle ear. Divers must push air into the ear through their Eustacean tubes to equalize this pressure.

Worked Example 63

pressure in the ear of a diver

How deep would you have to dive before the air in your middle ear would be compressed to 75% of its initial volume? Assume for the beginning that the temperature of the sea is constant as you dive.
Solution:
First we write down the pressure as a function of height h:

$$p = p_0 + \rho \cdot g \cdot h$$

where we take for p_0 the atmospheric pressure at height $h = 0$, ρ is the density of water at 20 degrees Celsius 998.23 kg/m^3, $g = 9.81$ ms^{-2}.

As the temperature is constant, it holds for both heights h

$$p_0 \cdot V_0 = (p_0 + \rho g h) \cdot V_e$$

Now solving for h using the fact that

$$V_e/V_0 = 0.75$$

yields

$$h = (0.75 * p_0 - p_0)/(\rho g)$$

Now, how far can the diver dive down before the membranes of his ear brake.

Solution:
As the result is negative, h determines the way he can dive down. h is given as roughly 2.6 m.

In 1809, the French chemist Joseph-Louis Gay-Lussac investigated the relationship between the Pressure of a gas and its temperature. Keeping a constant volume, the pressure of a gas sample is directly proportional to the temperature. Attention, the temperature is measured in Kelvin! The mathematical statement is as follows:

$$p_1/T_1 = p_2/T_2 = constant$$

$p_{1,2}$	: pressures (Pa)
$T_{1,2}$	: Temperatures (K)

That means, that pressure divided by temperature is a constant. On the other hand, if we plot pressure versus temperature, the graph crosses 0 pressure for $T = 0$ K = -273.15 °C as shown in the following figure. That point is called the ***absolute Zero***. That is where any motion of molecules, electrons or other particles stops.

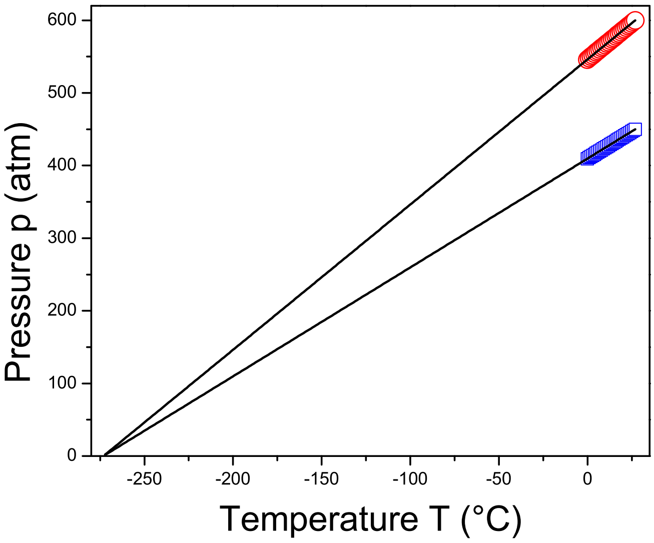


Figure 12.7: Pressure-temperature diagram for the ideal gas at constant volume.

Worked Example 64

Gay-Lussac

Suppose we have the following problem:
A gas cylinder containing explosive hydrogen gas has a pressure of 50 atm at a temperature of 300 K. The cylinder can withstand a pressure of 500 atm before it bursts, causing a building-flattening explosion. What is the maximum temperature the cylinder can withstand before bursting?
Solution: Let's rewrite this, identifying the variables:
A gas cylinder containing explosive hydrogen gas has a pressure of 50 atm (p_1) at a temperature of 300 K (T_1). The cylinder can withstand a pressure of 500 atm

(p_2) before it bursts, causing a building-flattening explosion. What is the maximum temperature the cylinder can withstand before bursting?
Plugging in the known variables into the expression for the Gay-Lussac law yields

$$T_2 = p_2/p_1 * T_1 = 500atm/50atm * 300K = 3000K$$

we find the answer to be 3000 K.

The law of combining volumes was interpreted by the Italian chemist Amedeo Avogadro in 1811, using what was then known as the Avogadro hypothesis. We would now properly refer to it as Avogadro's law:

Equal volumes of gases under the same conditions of temperature and pressure contain equal numbers of molecules.

This can be understood in the following. As in an ideal gas, all molecules are considered to be tiny particles with no spatial extension which collide elastically with each other. So, the kind of gas is irrelevant. Avogadro found that at room temperature, in atmospheric pressure the volume of a mol of a substance, i.e. $6.022 \cdot 10^{23}$ molecules or atoms, occupies the volume of 22.4 l.

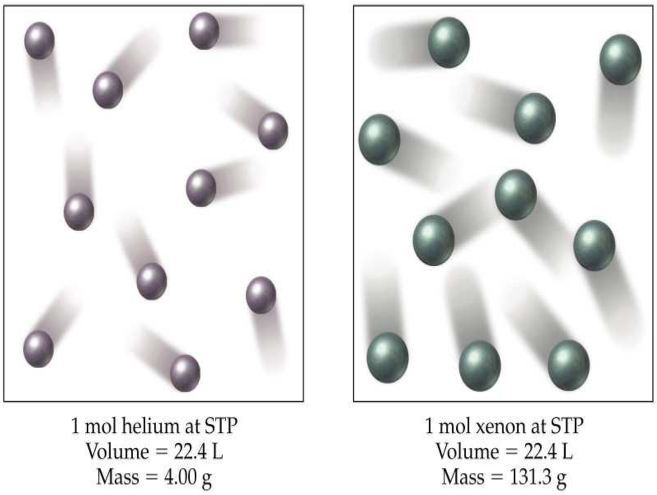


Figure 12.8: Two different gases occupying the same volume under the same circumstances.

Combination of the three empirical gas laws, described in the preceding three sections leads to the **Ideal Gas Law** which is usually written as:

$$p \cdot V = n \cdot R \cdot T$$

p	:	pressure (Pa)
V	:	Volume (m^3)
n	:	number of mols (mol)
R	:	gas konstant ($J/molK$)
T	:	temperature (K)

where p = pressure, V = volume, n = number of mols, T = kelvin temperature and R the ideal gas constant.

The ideal gas constant R in this equation is known as the universal gas constant. It arises from a combination of the proportionality constants in the three empirical gas laws. The universal gas constant has a value which depends only upon the units in which the pressure and volume are measured. The best available value of the universal gas constant is:

8.3143510 $\frac{J}{molK}$ or 8.3143510 $\frac{kPadm^3}{molK}$

Another value which is sometimes convenient is 0.08206 dm^3 atm/mol K. R is related to the Boltzmann-constant as:

$$R = N_0 \cdot k_B \tag{12.1}$$

where N_0 is the number of molecules in a mol of a substance, i.e. $6.022 \cdot 10^{23}$ and k_B is $1.308 \cdot 10^{-23}$ J/K is valid for one single particle.

This ideal gas equation is one of the most used equations in daily life, which we show in the following problem set:

Worked Example 65

ideal gas 1

A sample of 1.00 mol of oxygen at 50 °C and 98.6 kPa occupies what volume?
Solution:
We solve the ideal gas equation for the volume

$$V = \frac{nRT}{p}$$

and plug in the values $n = 1$, $T = 273.15 + 50$ K $= 323.15$ K and $p = 98.6 \cdot 10^3$ Pa, yielding for the volume $V = 0.0272$ m^3 = 27.2 dm^3.

This equation is often used to determine the molecular masses from gas data.

Worked Example 66

ideal gas 2

A liquid can be decomposed by electricity into two gases. In one experiment, one of the gases was collected. The sample had a mass of 1.090 g, a volume of 850 ml, a pressure of 746 torr, and a temperature of 25 °C. Calculate its molecular mass.
Solution:
To calculate the molecular mass we need the number of grams and the number of mols. We can get the number of grams directly from the information in the question. We can calculate the mols from the rest of the information and the ideal gas equation.

$$V = 850mL = 0.850L = 0.850dm^3$$
$$P = 746torr/760torr = 0.982atm$$
$$T = 25.0°C + 273.15 = 298.15K$$

$$pV = nRT$$
$$(0.982atm)(0.850L) = (n)(0.0821Latmmol - 1K - 1)(298.15K)$$
$$n = 0.0341mol$$

molecular mass = g/mol = 1.090 g/ 0.0341 mol = 31.96 g/mol. The gas is oxygen.

Or the equation can be comfortably used to design a gas temperature controller:

Worked Example 67

ideal gas 3

In a gas thermometer, the pressure needed to fix the volume of 0.20 g of Helium at 0.50 L is 113.3 kPa. What is the temperature?
Solution:
We transform first need to find the number of mols for Helium. Helium consists of 2 protons and 2 neutrons in the core (see later) and therefore has a molar volume of 4 g/mol. Therefore, we find

$$n = 0.20g/4g/mol = 0.05mol$$

plugging this into the ideal gas equation and solving for the temperature T we find:

$$T = \frac{pV}{nR} = \frac{113.3 \cdot 10^3 Pa \cdot 0.5 \cdot 10^{-3} m^3}{0.05mol \cdot 8.314J/molK} = 136.3K$$

The temperature is 136 Kelvin.

12.4.2 Kinetic theory of gasses

The results of several experiments can lead to a *scientific law*, which describes then all experiments performed. This is an empirical, that is based on experience only, approach to Physics. A law, however, only describes results; it does not explain why they have been obtained. Significantly stronger, a *theory* is a formulation which explains the results of experiments. A theory usually bases on *postulates*, that is a proposition that is accepted as true in order to provide a basis for logical reasoning. The most famous postulate in Physics is probably the one formulated by Walter Nernst which states that if one could reach absolute zero, all bodies would have the same entropy.

The kinetic-molecular theory of gases is a theory of great explanatory power. We shall see how it explains the ideal gas law, which includes the laws of Boyle and of Charles; Dalton's law of partial pressures; and the law of combining volumes.

The kinetic-molecular theory of gases can be stated as four postulates:

- A gas consists of particles (atoms or molecules) in continuous, random motion.
- Gas molecules influence each other only by collision; they exert no other forces on each other.
- All collisions between gas molecules are perfectly elastic; all kinetic energy is conserved.
- The average energy of translational motion of a gas particle is directly proportional to temperature.

In addition to the postulates above, it is assumed that the volumes of the particles are negligible as compared to container volume.

These postulates, which correspond to a physical model of a gas much like a group of billiard balls moving around on a billiard table, describe the behavior of an ideal gas. At room temperatures and pressures at or below normal atmospheric pressure, real gases seem to be accurately described by these postulates, and the consequences of this model correspond to the empirical gas laws in a quantitative way.

We define the average kinetic energy of translation E_t of a particle in a gas as

$$E_t = 1/2 \cdot mv^2 \tag{12.2}$$

where m is the mass of the particle with average velocity v. The forth postulate states that the average kinetic energy is a constant defining the temperature, i.e. we can formulate

$$E_t = 1/2 \cdot mv^2 = c \cdot T \tag{12.3}$$

where the temperature T is given in Kelvin and c is a constant, which has the same value for all gases. As we have 3 different directions of motion and each possible movement gives k_BT, we find for the energy of a particle in a gas as

$$E_t = 1/2 \cdot mv^2 = 3/2k_BT = 3/2\frac{R}{N_A}T \tag{12.4}$$

Hence, we can find an individual gas particle's speed rms = root mean square, which is the average square root of the speed of the individual particles (find u)

$$v_{rms} = \sqrt{\frac{3RT}{M_{mol}}} \tag{12.5}$$

where M_{mol} is the molar mass, i.e. the mass of the particle m times the Avogadro number N_A.

Worked Example 68

kinetic theory 1

Calculate the root-mean-square velocity of oxygen molecules at room temperature, 25 °C.
Solution:
Using

$$v_{rms} = \sqrt{3RT/M_{mol}} \ ,$$

the molar mass of molecular oxygen is 31.9998 g/mol; the molar gas constant has the value 8.3143 J/mol K, and the temperature is 298.15 K. Since the joule is the kg·m^2·s^{-2}, the molar mass must be expressed as 0.0319998 kg/mol. The root-mean-square velocity is then given by:

$$v_{rms} = \sqrt{3(8.3143)(298.15)/(0.0319998)} = 482.1 m/s$$

A speed of 482.1 m/s is 1726 km/h, much faster than a jetliner can fly and faster than most rifle bullets.

The very high speed of gas molecules under normal room conditions would indicate that a gas molecule would travel across a room almost instantly. In fact, gas molecules do not do so. If a small sample of the very odorous (and poisonous!) gas hydrogen sulfide is released in one corner of a room, our noses will not detect it in another corner of the room for several minutes unless the air is vigorously stirred by a mechanical fan. The slow diffusion of gas molecules which are moving very quickly occurs because the gas molecules travel only short distances in straight lines before they are deflected in a new direction by collision with other gas molecules.

The distance any single molecule travels between collisions will vary from very short to very long distances, but the average distance that a molecule travels between collisions in a gas can be calculated. This distance is called the *mean free path* l of the gas molecules. If the root-mean-square velocity is divided by the mean free path of the gas molecules, the result will be the number of collisions one molecule undergoes per second. This number is called the *collision frequency* Z_1 of the gas molecules.

The postulates of the kinetic-molecular theory of gases permit the calculation of the mean free path of gas molecules. The gas molecules are visualized as small hard spheres. A sphere of diameter d sweeps through a cylinder of cross-sectional area $\pi \cdot (d/2)^2$ and length v_{rms} each second, colliding with all molecules in the cylinder.

The radius of the end of the cylinder is d because two molecules will collide if their diameters overlap at all. This description of collisions with stationary gas molecules is not quite accurate, however, because the gas molecules are all moving relative to each other. Those relative velocities range between zero for two molecules moving in the same direction and $2v_{rms}$ for a head-on collision. The average relative velocity is that of a collision at right angles, which is $\sqrt{2}v_{rms}$. The total number of collisions per second per unit volume, Z_1, is

$$Z_1 = \pi d^2 \sqrt{2} v_{rms} \tag{12.6}$$

This total number of collisions must now be divided by the number of molecules which are present per unit volume. The number of gas molecules present per unit volume is found by rearrangement of the ideal gas law to $n/V = p/RT$ and use of Avogadro's number, $n = N/N_A$; thus $N/V = pN_A/RT$. This gives the mean free path of the gas molecules, l, as

$$(u_{rms}/Z_1)/(N/V) = l = RT/\pi d^2 p N_A \sqrt{2} \tag{12.7}$$

According to this expression, the mean free path of the molecules should get longer as the temperature increases; as the pressure decreases; and as the size of the molecules decreases.

Worked Example 69

mean free path

Calculate the length of the mean free path of oxygen molecules at room temperature, 25 °C, taking the molecular diameter of an oxygen molecule as 370 pm.
Solution:
Using the formula for mean free path given above and the value of the root-mean-square velocity u_{rms},

$$l = \frac{(8.3143kgm^2s^{-2}/Kmol)(298.15K)}{\pi(370\cdot10^{-12}m)2(101325kg/ms^2)(6.0225\cdot10^{23}mol^{-1})\sqrt{2}},$$

so l = 6.7·10^{-8} m = 67 nm.

The apparently slow diffusion of gas molecules takes place because the molecules travel only a very short distance before colliding. At room temperature and atmospheric pressure, oxygen molecules travel only (6.7·10^{-8} m)/(370·10^{-12} m) = 180 molecular diameters between collisions. The same thing can be pointed out using the collision frequency for a single molecule Z_1, which is the root-mean-square velocity divided by the mean free path:

$$Z_1 = \frac{\pi d^2 p N_A \sqrt{2}}{/RT} = v_{rms}/l \tag{12.8}$$

For oxygen at room temperature, each gas molecule collides with another every 0.13 nanoseconds (one nanosecond is 1.0·10^{-9} s), since the collision frequency is 7.2·10^{+9} collisions per second per molecule.

For an ideal gas, the number of molecules per unit volume is given using $pV = nRT$ and $n = N/N_A$ as

$$N/V = N_A p/RT \tag{12.9}$$

which for oxygen at 25 °C would be (6.022·10^{23} mol^{-1})(101325 kg/m s^2) / (8.3143 kg m^2/s^2 K mol)(298.15 K) or 2.46·10^{25} molecules/m^3. The number of collisions between two molecules in a volume, Z_{11}, would then be the product of the number of collisions each molecule makes times the number of molecules there are, $Z_1 N/V$, except that this would count each collision twice (since two molecules are involved in each one collision). The correct equation must be

$$Z_{11} = \frac{\pi d^2 p^2 N_A^2 \sqrt{2} v_{rms}}{2R^2T^2} \tag{12.10}$$

If the molecules present in the gas had different masses they would also have different speeds, so an average value of v_{rms} would be using a weighted average of the molar masses; the partial pressures of the different gases in the mixture would also be required. Although such calculations involve no new principles, they are beyond our scope.

12.4.3 Pressure of a gas

In the kinetic-molecular theory of gases, pressure is the force exerted against the wall of a container by the continual collision of molecules against it. From Newton's second law of motion, the force exerted on a wall by a single gas molecule of mass m and velocity v colliding with it is:

$$F = m \cdot a = m\frac{\Delta v}{\Delta t} \tag{12.11}$$

In the above equation, the change in a quantity is indicated by the symbol Δ, that means by changing the time t by a fraction, we change the velocity v by some other minimal amount. It is assumed that the molecule rebounds *elastically* and *no kinetic energy is lost* in a perpendicular collision, so Δv = v - (-v) = 2v (see figure below). If the molecule is moving perpendicular to the wall it will strike the opposite parallel wall, rebound, and return to strike the original wall again. If the length of the container or distance between the two walls is the path length l, then the time between two successive collisions on the same wall is Δt = 2l/v. The continuous force which the molecule moving perpendicular to the wall exerts is therefore

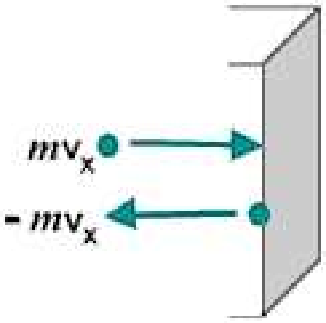


Figure 12.9: Change in momentum as a particle hits a wall.

$$F = m\frac{2v}{2l/v} = \frac{mv^2}{l} \tag{12.12}$$

The molecules in a sample of gas are not, of course, all moving perpendicularly to a wall, but the components of their actual movement can be considered to be along the three mutually perpendicular x, y, and z axes. If the number of molecules moving randomly, N, is large, then on the average one-third of them can be considered as exerting their force along each of the three perpendicular axes. The square of the average velocity along each axis, $v^2(x)$, $v^2(y)$, or $v^2(z)$, will be one-third of the square of the average total velocity v^2:

$$v^2(x) = v^2(y) = v^2(z) = v^2/3 \tag{12.13}$$

The average or mean of the square of the total velocity can replace the square of the perpendicular velocity, and so for a large number of molecules N,

$$F = (N/3)\frac{mv^2}{l} \tag{12.14}$$

Since pressure is force per unit area, and the area of one side of a cubic container must be l^2, the pressure p will be given by F/l^2 as:

$$p = (N/3)\frac{mv^2}{l^3} \tag{12.15}$$

This equation rearranges to

$$pV = N \cdot mv^2/3 \tag{12.16}$$

because volume V is the cube of the length l. The form of the ideal gas law given above shows the pressure-volume product is directly proportional to the mean-square velocity of the gas molecules. If the velocity of the molecules is a function only of the temperature, and we shall see in the next section that this is so, the kinetic-molecular theory gives a quantitative explanation of Boyle's law.

Worked Example 70

gas pressure

A square box contains He (Helium) at 25 °C. If the atoms are colliding with the walls perpendicularly (at 90°) at the rate of $4.0 \cdot 10^{22}$ times per second, calculate the force (in Newtons) and the pressure exerted on the wall per mol of He given that the area of the wall is 100 cm^2 and the speed of the atoms is 600 ms^{-1}.
Solution:
We use the equation 12.14 to calculate the force.

$$F = (N/3)\frac{mv^2}{l} = (N/3)mv\frac{v}{l}$$

The fraction v/l is the collision frequency $Z_1 = 0.6679$ s^{-1}. The product of $N \cdot Z_1$ is the number of molecules impinging on the wall per second. This induces for the force:

$$F = (N/3)mv\tau = 6.022 \cdot 1023/3 \cdot \frac{0.004g/mol}{6.022 \cdot 10^{23}} \cdot 600m/s \cdot 0.6679s^{-1}$$

yielding for the force $F = 0.534$ N. The pressure is the force per area:

$$p = F/A = 0.534N/0.01m^2 = 53.4Pa.$$

The calculated force is 0.534 N and the resulting pressure is 53.4 Pa.

12.4.4 Kinetic energy of molecules

In the following, we will make the connection between the kinetic theory and the ideal gas laws. We will find that the temperature is an important quantity which is the only intrinsic parameter entering in the kinetic energy of a gas.

We will consider an ensemble of molecules in a gas, where the molecules will be regarded as rigid large particles. We therefore neglect any vibrations or rotations in the molecule. Hence, making this assumption, Physics for a molecular gas is the same as for a single atom gas.

The square of the velocity is sometimes difficult to conceive, but an alternative statement can be given in terms of kinetic energy. The kinetic energy E_k of a single particle of mass m moving at velocity v is $mv^2/2$. For a large number of molecules N, the total kinetic energy E_k will depend on the mean-square velocity in the same way:

$$E_k = N \cdot mv^2/2 = n \cdot Mv^2/2 \tag{12.17}$$

The second form is on a molar basis, since $n = N/N_A$ and the molar mass $M = mN_A$ where N_A is Avogadro's number $6.022{\cdot}10^{23}$. The ideal gas law then appears in the form:

$$pV = 2E_k/3 \tag{12.18}$$

Compare $pV = nMv^2/2$. This statement that the pressure-volume product of an ideal gas is directly proportional to the total kinetic energy of the gas is also a statement of Boyle's law, since the total kinetic energy of an ideal gas depends only upon the temperature.

Comparison of the ideal gas law, $pV = nRT$, with the kinetic-molecular theory expression $pV = 2E_k/3$ derived in the previous section shows that the total kinetic energy of a collection of gas molecules is directly proportional to the absolute temperature of the gas. Equating the pV term of both equations gives

$$E_k = 3/2nRT \quad , \tag{12.19}$$

which rearranges to an explicit expression for temperature,

$$T = \frac{2}{3R}\frac{E_k}{n} = \frac{Mv^2}{3R} \tag{12.20}$$

We see that temperature is a function only of the mean kinetic energy E_k, the mean molecular velocity v, and the mean molar mass M.

Worked Example 71

mean velocity 1

Calculate the kinetic energy of 1 mol of nitrogen molecules at 300 K?
Solution:
Assume nitrogen behaves as an ideal gas, then

$$E_k = 3/2 \cdot RT = (3/2)8.3145J/(molK) \cdot 300K = 3742J/mol(or3.74kJ/mol)$$

At 300 K, any gas that behaves like an ideal gas has the same energy per mol.

As the absolute temperature decreases, the kinetic energy must decrease and thus the mean velocity of the molecules must decrease also. At $T = 0$, the absolute zero of temperature, all motion of gas molecules would cease and the pressure would then also be zero. No molecules would be moving. Experimentally, the absolute zero of temperature has never been attained, although modern experiments have extended to temperatures as low as 1 μK.

However, at low temperatures, the interactions between the particles becomes important and we enter a new regime of Quantum Mechanics, which considers molecules, single atoms or protons and electrons simultaneously as waves and as rigid particles. However, this would go too far.

Worked Example 72

mean velocity 2

If the translational rms. speed of the water vapor molecules (H_2O) in air is 648 m/s, what is the translational rms speed of the carbon dioxide molecules (CO_2) in the same air? Both gases are at the same temperature. And what is the temperature we measure?
Solution:
The molar mass of H_2O is

$$M_{H_2O} = 2 \cdot 1g/mol + 1 \cdot 16g/mol = 18g/mol$$

As the temperature is constant we can write

$$T = \frac{Mv^2}{3R} = \frac{0.018kg/mol \cdot (648m/s)^2}{3 \cdot 8.314J/mol \cdot K} = 303.0K = 29.9°C$$

Now we calculate the molar mass of CO_2

$$M_{CO_2} = 2 \cdot 16g/mol + 1 \cdot 12g/mol = 44g/mol$$

The rms velocity is again calculated with eq. 12.20

$$v_{CO_2} = \sqrt{3\frac{R \cdot T}{M_{CO_2}}} = \sqrt{\frac{3 \cdot 8.314J/molK \cdot 303.0K}{0.044kg/mol}} = 414.5m/s$$

The experiment was performed at 29.9 °C and the speed of the CO_2-molecules is 414.5 m/s, that is much slower than the water molecules as they are much heavier.

12.5 Temperature

Let us look back to the equation for the temperature of an ideal gas,

$$T = \frac{2}{3R}\frac{E_k}{n} \tag{12.21}$$

We can see that temperature is proportional to the average kinetic energy of a molecule in the gas. In other words temperature is a measure of how much energy is contained in an object – in hot things the atoms have a lot of kinetic energy, in cold things they have less. It may be surprising that ‘hot’ and ‘cold’ are really just words for how fast molecules or atoms are moving around, but it is true.

Definition: Temperature is a measure of the average kinetic energy of the particles in a body.

It should now be clear that heat is nothing more than energy on the move. It can be carried

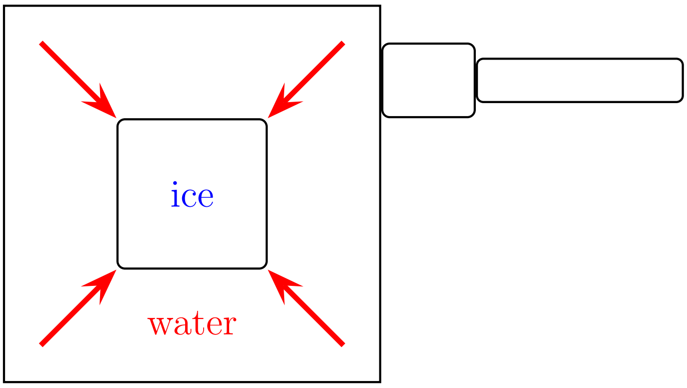


Figure 12.10: A heat flow diagram showing the heat flowing from the warmer water into the cooler ice cube.

by atoms, molecules or electromagnetic radiation but it is always just transport of energy. This is very important when we describe movement of heat as we will do in the following sections. 'Cold' is *not* a physical thing. It does not move from place to place, it is just the word for a lack of heat, just like dark is the word for an absence of light.

12.5.1 Thermal equilibrium

Now that we have defined the temperature of an isolated object (usually referred to as a body) we need to consider how heat will move between bodies at different temperatures. Let us take two bodies; A which has a fixed temperature and B whose temperature is allowed to change. If we allow heat to move between the two bodies we say they are in thermal contact.

First let us consider what happens if B is cooler than A. Remember – we have fixed the temperture of A so we need only worry about the temperature of B changing. An example of such a situation is an ice cube being dropped into a large pan of boiling water on a fire. The water temperature is fixed *i.e.* does not change, because the fire keeps it constant. It should be obvious that the ice cube will heat up and melt. In physical terms we say that the heat is flowing out of the (warmer) boiling water, into the (cooler) ice cube. This flow of heat into the ice cube causes it to warm up and melt. In fact the temperature of any cooler object in thermal contact with a warmer one will increase as heat from the warmer object flows into it.

The reverse would be true if B were warmer than A. We can now picture putting a small amount of warm water in to a freezer. If we come back in an hour or so the water will have cooled down and possibly frozen. In physical terms we say that the heat is flowing out of the (warmer) water, into the (cooler) air in the freezer. This flow of heat out of the ice cube in to the aircauses it to cool down and (eventually) freeze. Again, any warm object in thermal contact with a cooler one will cool down due to heat flowing out of it.

There is one special case which we have not yet discussed – what happens if A and B are at the same temperature? In this case B will neither warm up nor cool down, in fact, its temperature will remain constant. When two bodies are at the same temperature we say that they are in thermal equilibrium. Another way to express this is to say that two bodies are in thermal equilibrium if the particles within those bodies have the same average kinetic energies.

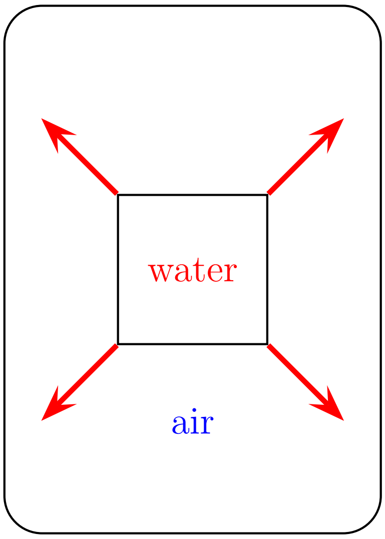


Figure 12.11: A heat flow diagram showing the heat flowing from the warmer water into the cooler air in the fridge.

Convince yourself that the last three paragraghs are correct before you continue. You should notice that heat always flows *from* the warmer object *to* the cooler object, *never* the other way around. Also, we never talk about coldness moving as it is not a real physical thing, only a lack of heat. Most importantly, it should be clear that the flow of heat between the two objects always attempts to bring them to the same temperature (or in other words, into thermal equilibrium). The logical conclusion of all this is that if two bodies are in thermal contact heat will flow from the hotter object to the cooler one until they are in thermal equilibrium (i.e. at the same temperature). We will see how to deal with this if the temperature of object A is not fixed in the section on heat capacities.

12.5.2 Temperature scales

Temperature scales are often confusing and even university level students can be tricked into using the wrong one. For most purposes in physics we do not use the familiar celcius (often innaccurately called centigrade) scale but the closely related absolute (or kelvin) scale – why? Let us think about the celcius scale now that we have defined temperature as a measure of the average kinteic energy of the atoms or molecules in a body.

A scale is a way of assigning a number to a physical quantity. Consider distance – using a ruler we can measure a distance and find its legngth. This legnth could be measured in metres, inches, or miles. The same is true of temperatures in that many different scales exist to measure them. Table 12.3 shows a few of these scales. Just like a ruler the scales have two defined points which fix the scale (consider the values at the beginning and end of the ruler e.g. $0cm$ and $15cm$). This is usually achieved by defining the temperature of some physical process, e.g. the freezing point of water.

Armed with our knowledge of temperture we can see that Celcius's scale has a big problem – it allows us to have a negative temperature.

Scale	Symbol	Definition
Fahrenheit	°F	Temperature at which an equal mixture of ice and salt melts = 0°F Temperature of blood = 96°F
Celcius	°C	Temperature at which water freezes = 0°C Temperature at which water boils = 100°C
Kelvin	K	Absolute zero is 0 K Triple point of water is 273.16 K

Table 12.3: The most important temperature scales.

We found that temperature is a measure of the average kinetic energy of the particles in a body. Therefore, a negative temperature suggests that that the particles have negative kinetic energy. This can not be true as kinetic energy can only be positive. Kelvin addressed this problem by redefining the zero of the scale. He realised that the coldest temperature you could achieve would be when the particles in a body were not moving at all. There is no way to cool something further than this as there is no more kinetic energy to remove from the body. This temperature is called *absolute zero*. Kelvin chose his scale so that $0K$ was the same as absolute zero and chose the size of his degree to be the same as one degree in the celcius scale.

Interesting Fact: Rankine did a similar thing to Kelvin but set his degree to be the same size as one degree fahrenheit. Unfortunately for him, almost everyone preferred Kelvin's absolute scale and the rankine scale is now hardly ever used!

It turns out that the freezing point of water, 0°C, is equal to 273.15 K. So, in order to convert from celcius to kelvin need to subtact 273.15.

Definition: $T(K) = T(°C) - 273.15$

12.5.3 Practical thermometers

It is often important to be able to determine an object's temperature precisely. This can be a challenge at very high or low tempertures or in inaccesible places. Consider a scientist who wishes to know how hot the magma in a volcano is. They are not going to be able to just lower a thermometer in to the magma as it will just melt as it reaches the superheated rock.

We will now look at some less extreme situations and show how a variety of thermometry techniques can be developed. Consider first the gas cylinder which we tried to explode in worked example 7 by heating it while sealed. We decided that we would need to heat it to around $3000K$ before it explodes. How can we check this experimentally? In a sealed gas cylinder the volume of the gas and the number of moles of gas remain constant as we heat, this is why we could use the Gay-Lussac law in example 7. The Gay-Lussac law tells us that pressure is directly proportional

to temperature for fixed volume and amount of gas. Therefore by mesuring the pressure in the cylinder (which can be done by fitting a pressure gauge to the top of it) we can indirectly work out the temperature.
This is similar to familiar alcohol or mercury thermometers. In these we use the fact that expansion of a liquid as it is heated is approximately proportional to temperature so we can use this expansion to as a measure of temperature. In fact, any thermometer you can imagine uses some physical property which varies with temperature to measure it indirectly.

12.5.4 Specific heat capacity

Conversion of macroscopic energy to microscopic kinetic energy thus tends to raise the temperature, while the reverse conversion lowers it. It is easy to show experimentally that the amount of heating needed to change the temperature of a body by some amount is proportional to the amount of matter in the body. Thus, it is natural to write

$$\Delta Q = MC\Delta T$$

(23.4)

where M is the mass of material, ΔQ is the amount of energy transferred to the material, and ΔT is the change of the material's temperature. The quantity C is called the specific heat of the material in question and is the amount of energy needed to raise the temperature of a unit mass of material one degree in temperature. C varies with the type of material. Values for common materials are given in table 22.2.

Table 22.2: Specific heats of common materials. Material C (J kg^{-1} K^{-1}) brass 385 glass 669 ice 2092 steel 448 methyl alcohol 2510 glycerine 2427 water 4184

12.5.5 Specific latent heat

It can be seen that the specific heat as defined above will be infinitely large for a phase change, where heat is transferred without any change in temperature. Thus, it is much more useful to define a quantity called latent heat, which is the amount of energy required to change the phase of a unit mass of a substance at the phase change temperature.

12.5.6 Internal energy

In thermodynamics, the internal energy is the energy of a system due to its temperature. The statement of first law refers to thermodynamic cycles. Using the concept of internal energy it is possible to state the first law for a non-cyclic process. Since the first law is another way of stating the conservation of energy, the energy of the system is the sum of the heat and work input, i.e., E = Q + W. Here E represents the heat energy of the system along with the kinetic energy and the potential energy (E = U + K.E. + P.E.) and is called the total internal energy of the system. This is the statement of the first law for non-cyclic processes.

For gases, the value of K.E. and P.E. is quite small, so the important internal energy function is U. In particular, since for an ideal gas the state can be specified using two variables, the state variable u is given by , where v is the specific volume and t is the temperature. Thus, by definition, , where cv is the specific heat at constant volume.

Internal energy of an Ideal gas

In the previous section, the internal energy of an ideal gas was shown to be a function of both the volume and temperature. Joule performed an experiment where a gas at high pressure inside a bath at the same temperature was allowed to expand into a larger volume.

picture required

In the above image, two vessels, labeled A and B, are immersed in an insulated tank containing water. A thermometer is used to measure the temperature of the water in the tank. The two vessels A and B are connected by a tube, the flow through which is controlled by a stop. Initially, A contains gas at high pressure, while B is nearly empty. The stop is removed so that the vessels are connected and the final temperature of the bath is noted.

The temperature of the bath was unchanged at the and of the process, showing that the internal energy of an ideal gas was the function of temperature alone. Thus Joule's law is stated as = 0.

12.5.7 First law of thermodynamics

We now address some questions of terminology. The use of the terms "heat" and "quantity of heat" to indicate the amount of microscopic kinetic energy inhabiting a body has long been out of favor due to their association with the discredited "caloric" theory of heat. Instead, we use the term internal energy to describe the amount of microscopic energy in a body. The word heat is most correctly used only as a verb, e. g., "to heat the house". Heat thus represents the transfer of internal energy from one body to another or conversion of some other form of energy to internal energy. Taking into account these definitions, we can express the idea of energy conservation in some material body by the equation

$$\Delta E = \Delta Q - \Delta W \quad \text{(first law of thermodynamics)}$$

where ΔE is the change in internal energy resulting from the addition of heat ΔQ to the body and the work ΔW done by the body on the outside world. This equation expresses the first law of thermodynamics. Note that the sign conventions are inconsistent as to the direction of energy flow. However, these conventions result from thinking about heat engines, i. e., machines which take in heat and put out macroscopic work. Examples of heat engines are steam engines, coal and nuclear power plants, the engine in your automobile, and the engines on jet aircraft.

12.6 Important Equations and Quantities

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
			or	

Table 12.4: Units used in **Electricity and Magnetism**

Chapter 13

Electrostatics

13.1 What is Electrostatics?

Electrostatics is the study of electric charge which is not moving *i.e. is static.*

13.2 Charge

All objects surrounding us (including people!) contain large amounts of electric charge. Charge can be **negative** or **positive** and is measured in units called **coulombs (C)**. Usually, objects contain the same amount of positive and negative charge so its effect is not noticeable and the object is called **electrically neutral**. However, if a small imbalance is created (i.e. there is a little bit more of one type of charge than the other on the object) then the object is said to be **electrically charged**.

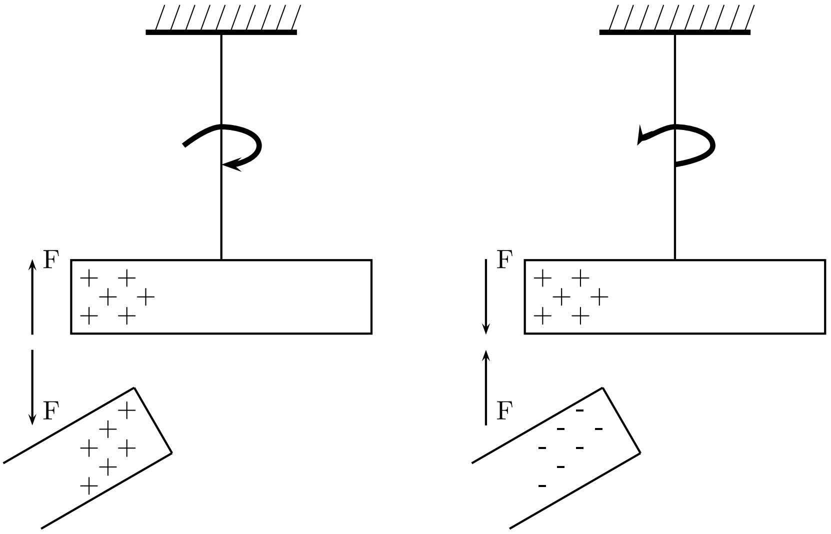
Some rather amusing examples of what happens when a person becomes charged are for example when you charge your hair by combing it with a plastic comb and it stands right up on end! Another example is when you walk fast over a nylon carpet and then touch a metal doorknob and give yourself a small shock (alternatively you can touch your friend and shock them!)

Charge has 3 further important properties:

- Charge is always **conserved**. Charge, just like energy, cannot be created or destroyed.
- Charge comes in discrete packets. The smallest unit of charge is that carried by one electron called the **elementary charge, e**, and by convention, it has a negative sign ($e = -1.6 \times 10^{-19}$C).
- Charged objects exert electrostatic forces on each other. **Like** charges **repel** and **unlike** charges **attract** each other.

Interesting Fact: The word ‘electron’ comes from the Greek word for amber! The ancient Greeks observed that if you rubbed a piece of amber, you could use it to pick up bits of straw. (Attractive electrostatic force!)

You can easily test the fact that like charges repel and unlike charges attract by doing a very simple experiment. Take a glass rod and rub it with a piece of silk, then hang it from its middle with a piece string so that it is free to move. If you then bring another glass rod which you have also charged in the same way next to it, you will see the rod on the string turn *away* from the rod in your hand i.e. it is **repelled**. If, however, you take a plastic rod, rub it with a piece of fur and then bring it close to the rod on the string, you will see the rod on the string turn *towards* the rod in your hand i.e. it is **attracted**!



What actually happens is that when you rub the glass with silk, tiny amounts of negative charge are transferred from the glass onto the silk, which causes the glass to have less negative charge than positive charge, making it **positively charged**. When you rub the plastic rod with the fur, you transfer tiny amounts of negative charge onto the rod and so it has more negative charge than positive charge on it, making it **negatively charged**.

Conductors and Insulators

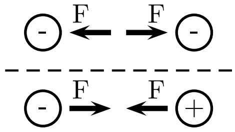
Some materials allow charge carriers to move relatively freely through them (e.g. most metals, tap water, the human body) and these materials are called **conductors**. Other materials, which do not allow the charge carriers to move through them (e.g. plastic, glass), are called **non-conductors** or **insulators**.

Aside: As mentioned above, the basic unit of charge, namely the elementary charge, e, is carried by the electron. In a conducting material (e.g. copper), when the atoms bond to form the material, some of the outermost, loosely bound electrons become detached from the individual atoms and so become free to move around. The charge carried by these electrons can move around in the material! In insulators, there are very few, if any, free electrons and so the charge cannot move around in the material.

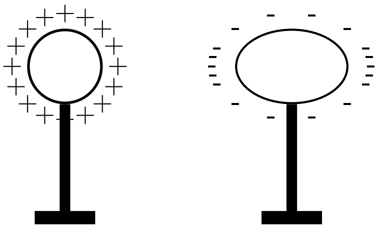
If an excess of charge is put onto an insulator, it will stay where it is put and there will be a concentration of charge in that area on the object. However, if an excess of charge is put onto a conductor, the charges of like sign will repel each other and spread out over the surface of the object. When two conductors are made to touch, the total charge on them is shared between the two. If the two conductors are identical, then each conductor will be left with half of the total charge.

13.3 Electrostatic Force

As we now know, charged objects exert a force on one another. If the charges are at rest then this force between them is known as the **electrostatic force**. An interesting characteristic of the electrostatic force is that it can be either attractive or repulsive, unlike the gravitational force which is only ever attractive. The relative charges on the two objects is what determines whether the force between the charged objects is attractive or repulsive. If the objects have opposite charges they attract each other, while if their charges are similar they repel each other (e.g. two metal balls which are negatively charged will repel each other, while a positively charged ball and negatively charged ball will attract one another).



It is this force that determines the arrangement of charge on the surface of conductors. When we place a charge on a spherical conductor the repulsive forces between the individual like charges cause them to spread uniformly over the surface of the sphere. However, for conductors with non-regular shape there is a concentration of charge near the point or points of the object.



This collection of charge can actually allow charge to leak off the conductor if the point is sharp enough. It is for this reason that buildings often have a lightning rod on the roof to remove any charge the building has collected. This minimises the possibility of the building being struck by lightning.

This "spreading out" of charge would not occur if we were to place the charge on an insulator since charge cannot move in insulators.

13.3.1 Coulomb's Law

The behaviour of the electrostatic force was studied in detail by Charles Coulomb around 1784. Through his observations he was able to show that the electrostatic force between two point-like charges is inversely proportional to the square of the distance between the objects. He also discovered that the force is proportional to the product of the charges on the two objects.

$$F \propto \frac{Q_1 Q_2}{r^2},$$

where Q_1 is the charge on the one point-like object, Q_2 is the charge on the second, and r is the distance between the two.

The magnitude of the electrostatic force between two point-like charges is given by *Coulomb's Law*:

$$F = k\frac{Q_1 Q_2}{r^2} \tag{13.1}$$

and the proportionality constant k is called the *electrostatic constant*. We will use the value

$$k = 8.99 \times 10^9 \text{N} \cdot \text{m}^2/\text{C}^2.$$

The value of the electrostatic constant is known to a very high precision (9 decimal places). Not many physical constants are known to as high a degree of accuracy as k.

> *Aside:* Notice how similar Coulomb's Law is to the form of Newton's Universal Law of Gravitation between two point-like particles:
>
> $$F_G = G\frac{m_1 m_2}{r^2},$$
>
> where m_1 and m_2 are the masses of the two particles, r is the distance between them, and G is the gravitational constant. It is very interesting that Coulomb's Law has been shown to be correct no matter how small the distance, nor how large the charge: for example it still applies inside the atom (over distances smaller than 10^{-10}m).

Let's run through a simple example of electrostatic forces.

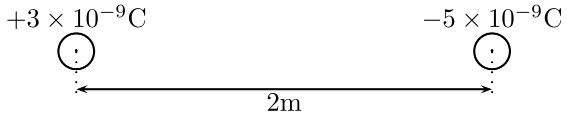
Worked Example 73

Coulomb's Law I

Question: Two point-like charges carrying charges of $+3 \times 10^{-9}$C and -5×10^{-9}C are 2m apart. Determine the magnitude of the force between them and state whether it is attractive or repulsive. **Answer:**

Step 1 : (NOTE TO SELF: step is deprecated, use westep instead.)

First draw the situation:



Step 2 : (NOTE TO SELF: step is deprecated, use westep instead.)

Is everything in the correct units? Yes, charges are in coulombs [C] and distances in meters [m].

Step 3 : (NOTE TO SELF: step is deprecated, use westep instead.)
Determine the magnitude of the force: Using Coulomb's Law we have

$$
\begin{aligned}
F &= k\frac{Q_1Q_2}{r^2} \\
&= (8.99 \times 10^9 \mathrm{N}\cdot\mathrm{m}^2/\mathrm{C}^2)\frac{(+3\times10^{-9}\mathrm{C})(-5\times10^{-9}\mathrm{C})}{(2\mathrm{m})^2} \\
&= -3.37\times10^{-8}\mathrm{N}
\end{aligned}
$$

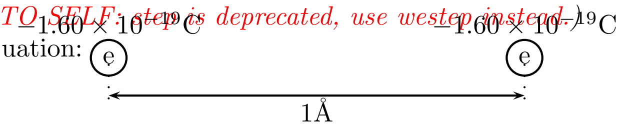
Thus the *magnitude* of the force is 3.37×10^{-8}N. The minus sign is a result of the two point charges having opposite signs.
Step 4 : (NOTE TO SELF: step is deprecated, use westep instead.)
Is the force attractive or repulsive? Well, since the two charges are oppositely charged, the force is *attractive*. We can also conclude this from the fact that Coulomb's Law gives a negative value for the force.

Next is another example that demonstrates the difference in magnitude between the gravitational force and the electrostatic force.

Worked Example 74

Coulomb's Law II

Question: Determine the electrostatic force and gravitational force between two electrons 1Åapart (i.e. the forces felt inside an atom)
Answer:
Step 1 : (NOTE TO SELF: step is deprecated, use westep instead.)
First draw the situation:



Step 2 : (NOTE TO SELF: step is deprecated, use westep instead.)
Get everything into S.I. units: The charge on an electron is -1.60×10^{-19}C, the mass of an electron is 9.11×10^{-31}kg, and 1Å=1×10^{-10}m
Step 3 : (NOTE TO SELF: step is deprecated, use westep instead.)
Calculate the electrostatic force using Coulomb's Law:

$$
\begin{aligned}
F_E &= k\frac{Q_1Q_2}{r^2} = k\frac{e\cdot e}{1\text{Å}^2} \\
&= (8.99\times10^9\mathrm{N}\cdot\mathrm{m}^2/\mathrm{C}^2)\frac{(-1.60\times10^{-19}\mathrm{C})(-1.60\times10^{-19}\mathrm{C})}{(10^{-10}\mathrm{m})^2} \\
&= 2.30\times10^{-8}\mathrm{N}
\end{aligned}
$$

Hence the *magnitude* of the electrostatic force between the electrons is 2.30×10^{-8}N. (Note that the electrons carry like charge and from this we know the force must be repulsive. Another way to see this is that the force is positive and thus repulsive.)

Step 4 : (NOTE TO SELF: step is deprecated, use westep instead.)
Calculate the gravitational force:

$$\begin{aligned}F_E &= G\frac{m_1m_2}{r^2} = G\frac{m_e \cdot m_e}{(1\mathring{A})^2}\\ &= (6.67\times10^{-11}\mathrm{N\cdot m^2/kg^2})\frac{(9.11\times10^{-31}\mathrm{C})(9.11\times10^{-31}\mathrm{kg})}{(10^{-10}\mathrm{m})^2}\\ &= 5.54\times10^{-51}\mathrm{N}\end{aligned}$$

The magnitude of the gravitational force between the electrons is 5.54×10^{-51}N

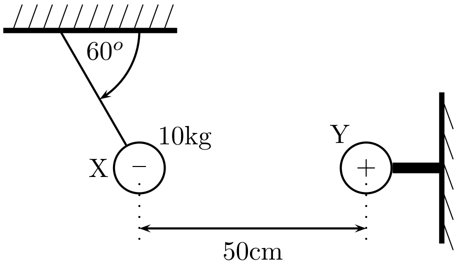
Notice that the gravitational force between the electrons is much smaller than the electrostatic force. For this reason, the gravitational force is usually neglected when determining the force between two charged objects.

We mentioned above that charge placed on a spherical conductor spreads evenly along the surface. As a result, if we are far enough from the charged sphere, electrostatically, it behaves as a point-like charge. Thus we can treat spherical conductors (e.g. metallic balls) as point-like charges, with all the charge acting at the centre.

Worked Example 75

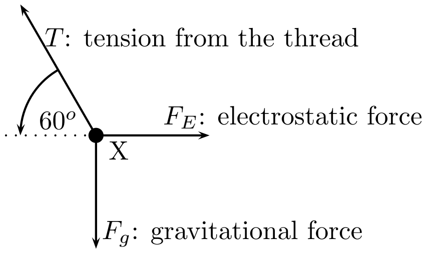
Coulomb's Law: Challenge Question

Question: In the picture below, X is a small negatively charged sphere with a mass of 10kg. It is suspended from the roof by an insulating rope which makes an angle of 60^o with the roof. Y is a small positively charged sphere which has the same magnitude of charge as X. Y is fixed to the wall by means of an insulating bracket. Assuming the system is in equilibrium, what is the magnitude of the charge on X?



Answer:
How are we going to determine the charge on X? Well, if we know the force between X and Y we can use Coulomb's Law to determine their charges as we know the distance between them. So, firstly, we need to determine the magnitude of the electrostatic force between X and Y.
Step 1 : (NOTE TO SELF: step is deprecated, use westep instead.)
Is everything in S.I. units? The distance between X and Y is 50cm = 0.5m, and the mass of X is 10kg.

Step 2 : (NOTE TO SELF: step is deprecated, use westep instead.)
Draw the forces on X (with directions) and label.



T: tension from the thread
60^o
F_E: electrostatic force
X
F_g: gravitational force

Step 3 : (NOTE TO SELF: step is deprecated, use westep instead.)
Determine the magnitude of the electrostatic force (F_E). Since nothing is moving (system is in equlibrium) the vertical and horizonal components of the forces must cancel. Thus

$$F_E = T\sin(60^o); \qquad F_g = T\sin(60^o).$$

The only force we know is the gravitational force $F_g = mg$. Now we can calculate the magnitude of T from above:

$$T = \frac{F_g}{\sin(60^o)} = \frac{(10\mathrm{kg})(10\mathrm{m/s}^2)}{\sin(60^o)} = 1155\mathrm{N}.$$

Which means that F_E is:

$$F_E = T\cos(60^o) = 1154\mathrm{N}\cdot\cos(60^\circ) = 577.5\mathrm{N}$$

Step 4 : (NOTE TO SELF: step is deprecated, use westep instead.)
Now that we know the magnitude of the electrostatic force between X and Y, we can calculate their charges using Coulomb's Law. Don't forget that the magnitudes of the charges on X and Y are the same: $|Q_\mathrm{X}| = |Q_\mathrm{Y}|$. The magnitude of the electrostatic force is

$$\begin{aligned} F_E &= k\frac{|Q_\mathrm{X}Q_\mathrm{Y}|}{r^2} = k\frac{Q_\mathrm{X}^2}{r^2} \\ |Q_\mathrm{X}| &= \sqrt{\frac{F_E r^2}{k}} \\ &= \sqrt{\frac{(577.5\mathrm{N})(0.5\mathrm{m})^2}{8.99\times 10^9\mathrm{N}\cdot\mathrm{m}^2/\mathrm{C}^2}} \\ &= 1.27\times 10^{-4}\mathrm{C} \end{aligned}$$

Thus the charge on X is -1.27×10^{-4}C

13.4 Electric Fields

We have learnt that objects that carry charge feel forces from all other charged objects. It is useful to determine what the effect of a charge would be at every point surrounding it. To do this we need some sort of reference.

We know that the force that one charge feels due to another depends on both charges (Q_1 and Q_2). How then can we talk about forces if we only have one charge? The solution to this dilemna is to introduce a *test charge.* We then determine the force that would be exerted on it if we placed it at a certain location. If we do this for every point surrounding a charge we know what would happen if we put a test charge at any location.

This map of what would happen at any point we call a field map. It is a map of the electric field *due to* a charge. It tells us how large the force on a test charge would be and in what direction the force would be.

Our map consists of the lines that tell us how the test charge would move if it were placed there.

13.4.1 Test Charge

This is the key to mapping out an electric field. The equation for the force between two electric charges has been shown earlier and is:

$$F = k\frac{Q_1 Q_2}{r^2}. \tag{13.2}$$

If we want to map the field for Q_1 then we need to know exactly what would happen if we put Q_2 at every point around Q_1. But this obviously depends on the value of Q_2. This is a time when we need to agree on a *convention.* What should Q_2 be when we make the map? By convention we choose $Q_2 = +1C$.

This means that if we want to work out the effects on any other charge we only have to multiply the result for the test charge by the magnitude of the new charge.

The electric field strength is then just the force per unit of charge and has the same magnitude and direction as the force on our test charge but has different units:

$$E = k\frac{Q_1}{r^2} \tag{13.3}$$

The electric field is the force per unit of charge and hence has units of newtons per coulomb [N/C].

So to get the force the electric field exerts we use:

$$F = EQ \tag{13.4}$$

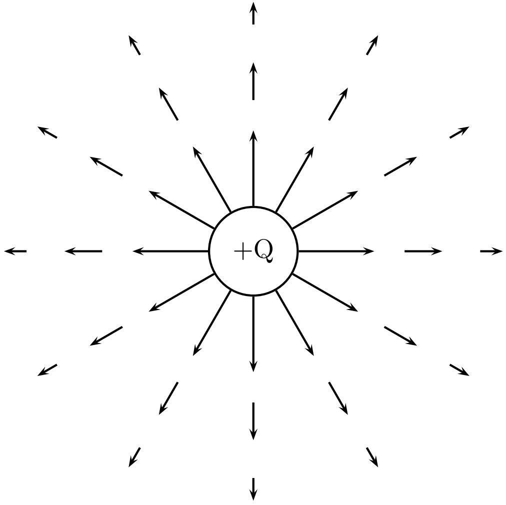
Notice we are just multiplying the electric field magnitude by the magnitude of the charge it is acting on.

13.4.2 What do field maps look like?

The maps depend very much on the charge or charges that the map is being made for. We will start off with the simplest possible case. Take a single positive charge with no other charges around it. First, we will look at what effects it would have on a test charge at a number of points.

Positive Charge Acting on Test Charge

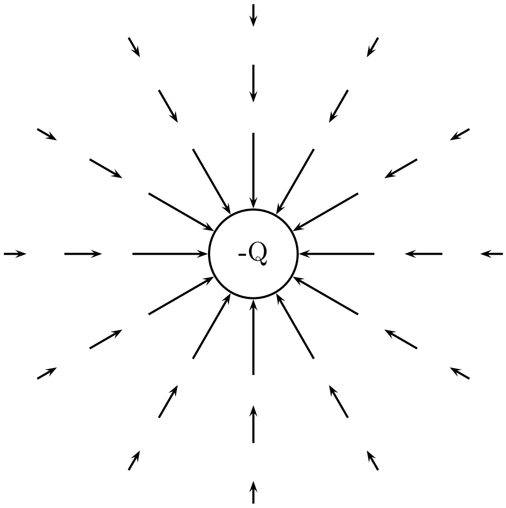
At each point we calculate the force on a test charge, q, and represent this force by a vector.



We can see that at every point the positive test charge, q, would experience a force pushing it away from the charge, Q. This is because both charges are positive and so they repel. Also notice that at points further away the vectors are shorter. That is because the force is smaller if you are further away.

If the charge were negative we would have the following result.

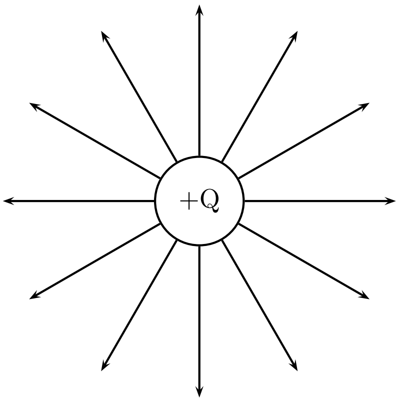
Negative Charge Acting on Test Charge



Notice that it is **almost** identical to the positive charge case. This is important – the arrows are the same length because the magnitude of the charge is the same and so is the magnitude of the test charge. Thus the **magnitude** of the force is the same. The arrows point in the opposite direction because the charges now have opposite sign and so the test charge is **attracted** to the charge.

Now, to make things simpler, we draw continuous lines showing the path that the test charge would travel. This means we don't have to work out the magnitude of the force at many different points.

Electric Field Map due to a Positive Charge



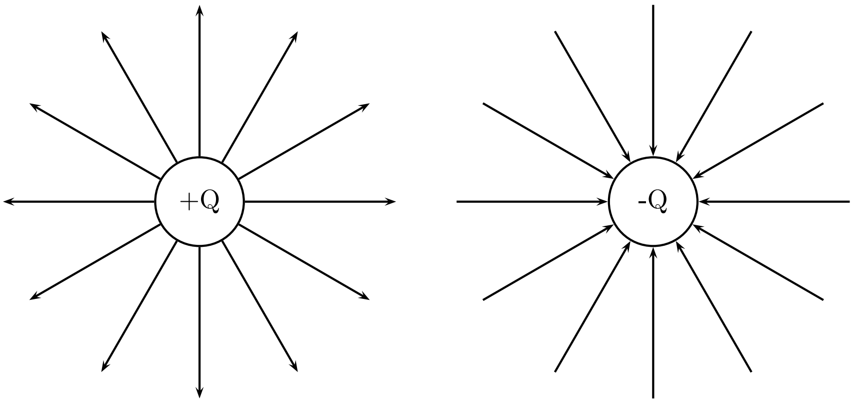
Some important points to remember about electric fields:

- There is an electric field at **every point** in space surrounding a charge.
- Field lines are merely a **representation** – they are not real. When we draw them, we just pick convenient places to indicate the field in space.
- Field lines always start at a **right-angle** (90^o) to the charged object causing the field.
- Field lines **never** cross!

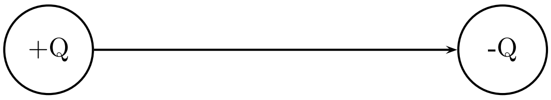
13.4.3 Combined Charge Distributions

We look at the field of a positive charge and a negative charge placed next to each other. The net resulting field would be the addition of the fields from each of the charges. To start off with let us sketch the field maps for each of the charges as though it were in isolation.

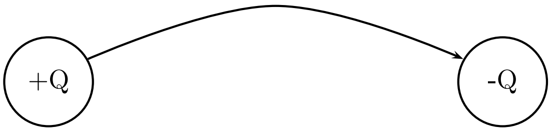
Electric Field of Negative and Positive Charge in Isolation



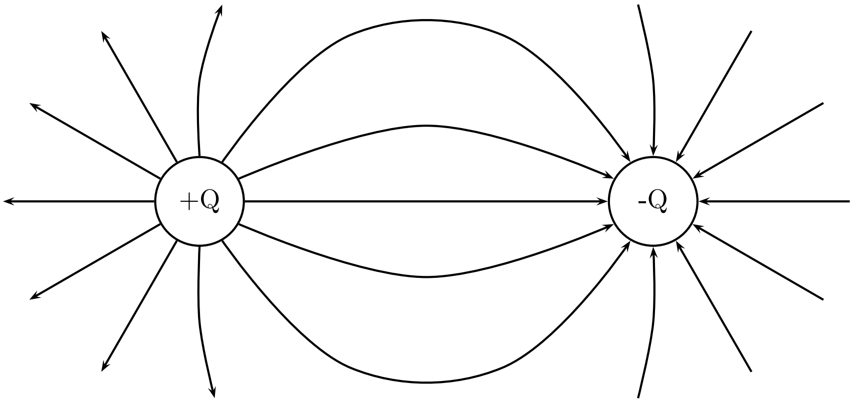
Notice that a test charge starting off directly between the two would be pushed away from the positive charge and pulled towards the negative charge in a straight line. The path it would follow would be a straight line between the charges.



Now let's consider a test charge starting off a bit higher than directly between the charges. If it starts closer to the positive charge the force it feels from the positive charge is greater, but the negative charge does attract it, so it would move away from the positive charge with a tiny force attracting it towards the negative charge. As it gets further from the positive charge the force from the negative and positive charges change and they are equal in magnitude at equal distances from the charges. After that point the negative charge starts to exert a stronger force on the test charge. This means that the test charge moves towards the negative charge with only a small force away from the positive charge.

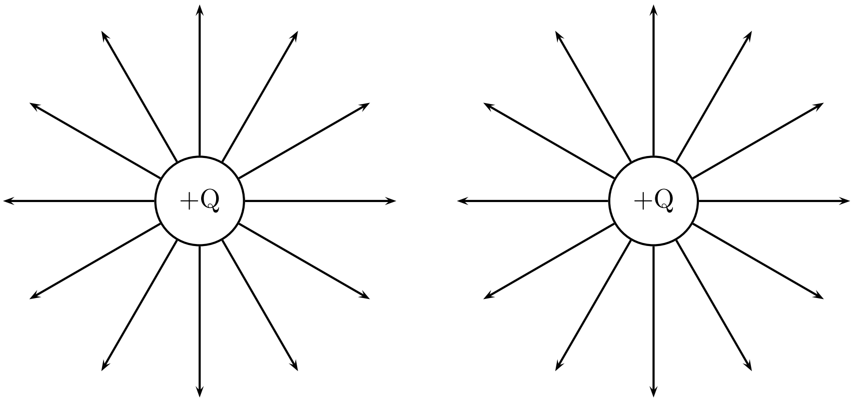


Now we can fill in the other lines quite easily using the same ideas. The resulting field map is:

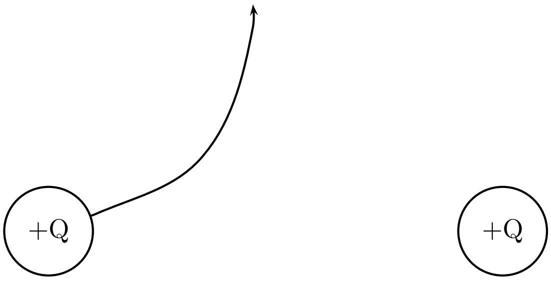


Two Like Charges I: The Positive Case

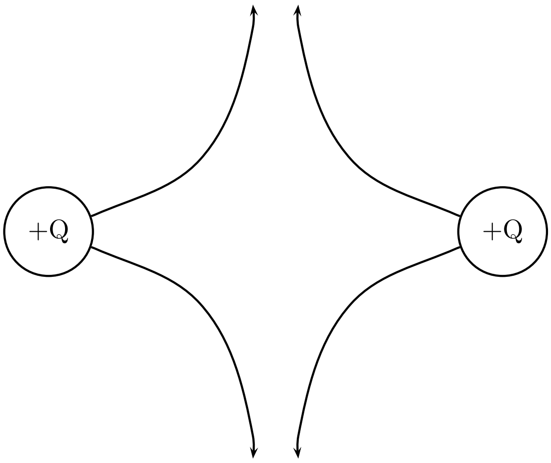
For the case of two positive charges things look a little different. We can't just turn the arrows around the way we did before. In this case the test charge is repelled by both charges. This tells us that a test charge will never cross half way because the force of repulsion from both charges will be equal in magnitude.



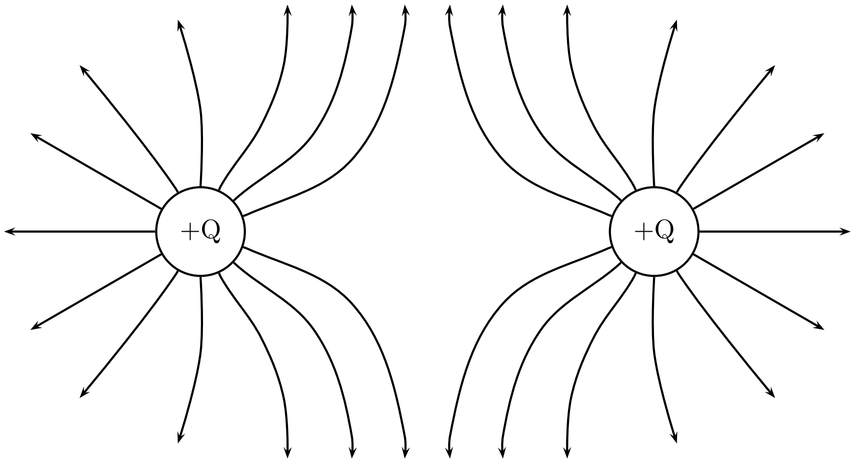
The field directly between the charges cancels out in the middle. The force has equal magnitude and opposite direction. Interesting things happen when we look at test charges that are not on a line directly between the two.



We know that a charge the same distance below the middle will experience a force along a reflected line, because the problem is symmetric (i.e. if we flipped vertically it would look the same). This is also true in the horizontal direction. So we use this fact to easily draw in the next four lines.

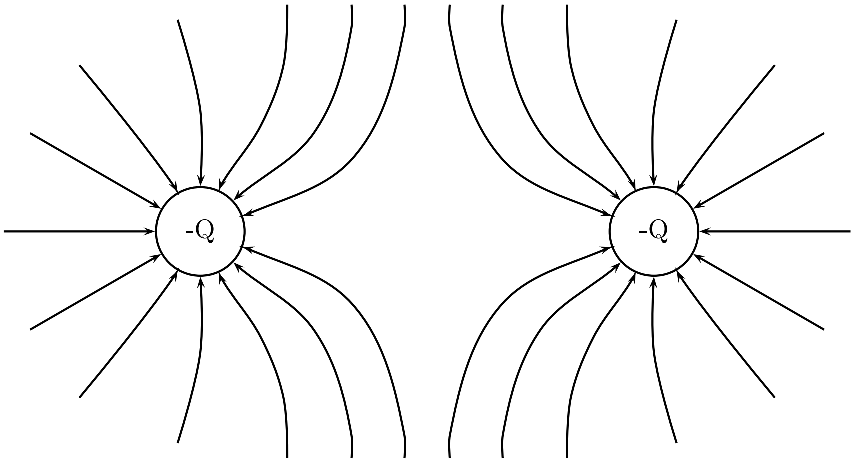


Working through a number of possible starting points for the test charge we can show the electric field map to be:



Two Like Charges II: The Negative Case

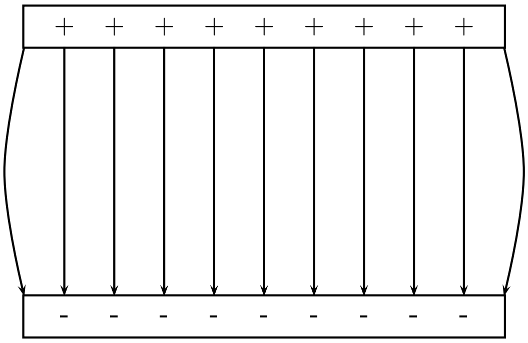
We can use the fact that the direction of the force is reversed for a test charge if you change the sign of the charge that is influencing it. If we change to the case where both charges are negative we get the following result:



13.4.4 Parallel plates

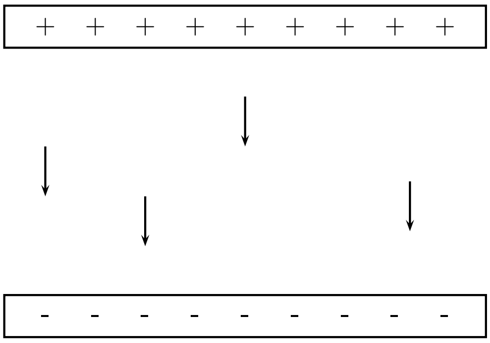
One very important example of electric fields which is used extensively is the electric field between two charged parallel plates. In this situation the electric field is constant. This is used for many practical purposes and later we will explain how Millikan used it to measure the charge on the electron.

Field Map for Oppositely Charged Parallel Plates



This means that the force that a test charge would feel at any point between the plates would be identical in magnitude and direction. The fields on the edges exhibit fringe effects, *i.e. they bulge outwards*. This is because a test charge placed here would feel the effects of charges only on one side (either left or right depending on which side it is placed). Test charges placed in the middle experience the effects of charges on both sides so they balance the components in the horizontal direction. This isn't the case on the edges.

The Force on a Test Charge between Oppositely Charged Parallel Plates



13.4.5 What about the Strength of the Electric Field?

When we started making field maps we drew arrows to indicate the strength of the field and the direction. When we moved to lines you might have asked "Did we forget about the field strength?". We did not. Consider the case for a single positive charge again:

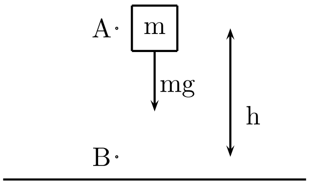
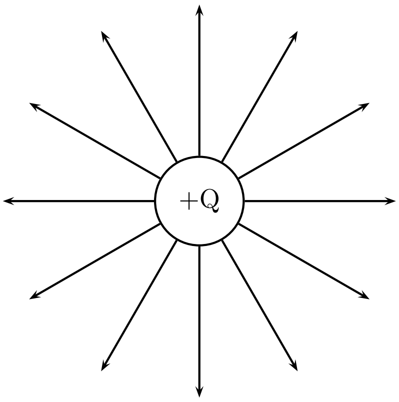


Figure 13.1: A mass under the influence of a gravitational field.



Notice that as you move further away from the charge the field lines become more spread out. In field map diagrams the closer field lines are together the stronger the field. This brings us to an interesting case. What is the electric field like if the object that is charged has an irregular shape.

13.5 Electrical Potential

13.5.1 Work Done and Energy Transfer in a Field

When a charged particle moves in an electric field work is done and energy transfers take place. This is exactly analogous to the case when a mass moves in a gravitational field such as that set up by any massive object.

Work done by a field

Gravitational Case
A mass held at a height h above the ground has **gravitational potential energy** since, if released, it will fall under the action of the gravitational field. Once released, in the absence of friction, only the force of gravity acts on the mass and the mass accelerates in the direction of the force (towards the earth's centre). In this way, work is done by the field. When the mass

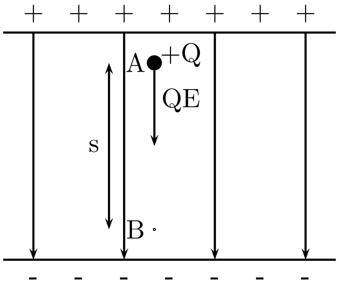


Figure 13.2: A charged particle under the influence of an electric field.

falls a distance h (from point A to B), the work done is,

$$\begin{aligned} W &= Fs \\ &= mgh \end{aligned}$$

Energy is **conserved!**

In falling, the mass loses gravitational potential energy and gains kinetic energy. The work done by the field is equal to the energy transferred,

$$W = \text{Gain in } E_k = \text{Loss in } E_p \qquad \text{(a falling mass)}$$

Electrical Case
A charge in an electric field has **electrical potential energy** since, if released, it will move under the action of the electric field. When released, in the absence of friction, only the electric force acts on the charge and the charge accelerates in the direction of the force (for positive charges the force and acceleration are in the direction of the electric field, while negative charges experience a force and acceleration in the opposite direction to the electric field.) Consider a positive charge $+Q$ placed in the uniform electric field between oppositely charged parallel plates. The positive charge will be repelled by the positive plate and attracted by the negative plate (i.e. it will move in the direction of the electric field lines). In this way, work is done by the field. In moving the charge a distance s in the electric field, the work done is,

$$\begin{aligned} W &= Fs \\ &= QEs \qquad \text{since } E = \frac{F}{Q}. \end{aligned}$$

In the process of moving, the charge loses electrical potential energy and gains kinetic energy. The work done by the field is equal to the energy transferred,

$$W = \text{Gain in } E_k = \text{Loss in } E_p \qquad \text{(charge moving under the influence of an electric field)}$$

Work done by us

Gravitational Case
In order to return the mass m in Fig.13.1 to its original position (i.e. lift it a distance h from B back to A) we have to apply a force mg to balance the force of gravity. An amount of work mgh is done by the lifter. In the process, the mass gains gravitational potential energy,

$$mgh = \text{Gain in } E_p \qquad \text{(lifting a mass)}$$

Electrical Case
In order to return the charge in Fig.13.2 to its original position (i.e. from B back to A) we have to exert a force QE on the charge to balance the force exerted on it by the electric field. An amount of work QEs is done by us. In the process, the charge gains electrical potential energy,

Energy is **conserved!**

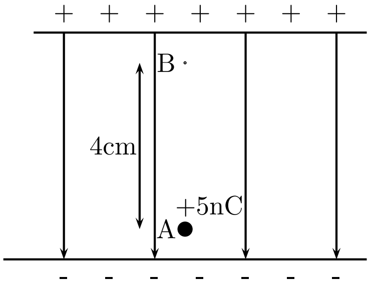
$$QEs \quad = \quad \text{Gain in } E_p \qquad \text{(charge moved against an electric field)}$$

In summary, **when an object moves under the influence of a field, the field does work and potential energy is transferred into kinetic energy. Potential energy is lost, while kinetic energy is gained. When an object is moved against a field we have to do work and the object gains potential energy.**

Worked Example 76

Work done and energy transfers in a field

Question: A charge of +5nC is moved a distance of 4 cm against a uniform electric field of magnitude 2×10^{12}N.C^{-1} from A to B.



(a) Calculate the work done in moving the charge from A to B.
(b) The charge is now released and returns to A. Calculate the kinetic energy of the charge at A.

Answer:
(a)
Step 1 : (NOTE TO SELF: step is deprecated, use westep instead.)
We are given the values of the charge, the field and the distance the charge must move. All are in the correct units.
Step 2 : (NOTE TO SELF: step is deprecated, use westep instead.)
Since the charge is positive we have to do work to move it from A to B (since this is against the field). This work is given by,

$$\begin{aligned} W &= QEs \\ &= (5 \times 10^{-9})(2 \times 10^{12})(0.04) \\ &= 400 \text{ J} \end{aligned}$$

(b)

Step 3 : (NOTE TO SELF: step is deprecated, use westep instead.)
When released the charge moves under the influence of the electric field and returns to A. Work is now done by the field and the work done is equal to the kinetic energy gained,

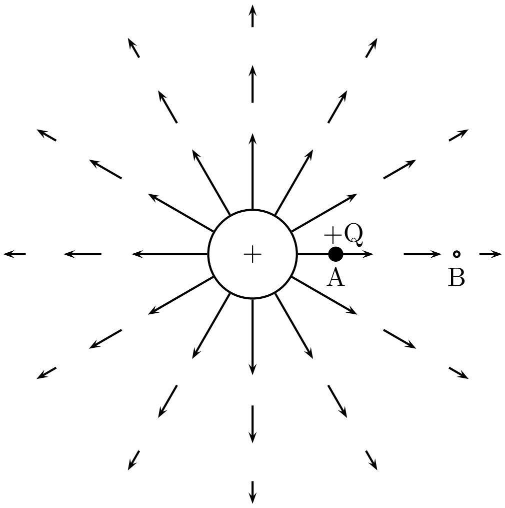
$$\begin{aligned}\text{Gain in } E_k &= QEs \\ &= (5 \times 10^{-9})(2 \times 10^{12})(0.04) \\ &= 400 \text{ J}\end{aligned}$$

Since the charge started at rest, the gain in kinetic energy is the final kinetic energy,

$$E_k^{\text{at A}} = 400 \text{ J}$$

13.5.2 Electrical Potential Difference

Consider a positive test charge $+Q$ placed at A in the electric field of another positive point charge.



The test charge moves towards B under the influence of the electric field of the other charge. In the process the test charge loses electrical potential energy and gains kinetic energy. Thus, at A, the test charge has more potential energy than at B – **A is said to have a higher electrical potential than B**. The potential energy of a charge at a point in a field is defined as the work required to move that charge from infinity to that point.

The **potential difference between two points** in an electric field is defined as the **work required to move a unit positive test charge from the point of lower potential to that of higher potential**. If an amount of work W is required to move a charge Q from one point to another, then the potential difference between the two points is given by,

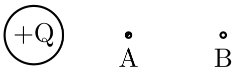
$$V = \frac{W}{Q} \qquad \text{unit : J.C}^{-1} \text{ or V (the volt)}$$

From this equation it follows that one volt is the potential difference between two points in an electric field if one joule of work is done in moving one coulomb of charge from the one point to the other.

Worked Example 77

Potential difference

Question:
A positively charged object Q is placed as shown in the sketch. The potential difference between two points A and B is 4×10^{-4} V.



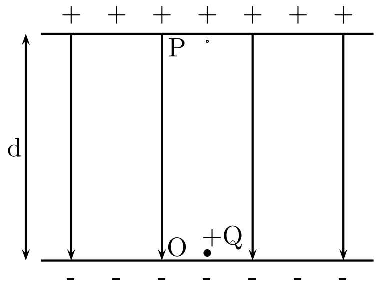
(a) Calculate the change in electrical potential energy of a +2nC charge when it moves from A to B.
(b) Which point, A or B, is at the higher electrical potential? Explain.
(c) If this charge were replaced with another of charge -2nC, in what way would its change in energy be affected?

Answer:
(a) The electrical potential energy of the positive charge decreases as it moves from A to B since it is moving in the direction of the electric field produced by the object Q. This loss in potential energy is equal to the work done by the field,

$$\begin{aligned}\text{Loss in Electrical Potential Energy} &= W \\ &= VQ \qquad (\text{Since } V = \frac{W}{Q}) \\ &= (4 \times 10^{-4})(2 \times 10^{-9}) \\ &= 8 \times 10^{-13}\ \text{J}\end{aligned}$$

(b) Point A is at the higher electrical potential since work is required by us to move a positive test charge from B to A.
(c) If the charge is replaced by one of negative charge, the electrical potential energy of the charge will increase in moving from A to B (in this case we would have to do work on the charge).

As an example consider the electric field between two oppositely charged parallel plates a distance d apart maintained at a potential difference V.



This electric field is uniform so that a charge placed anywhere between the plates will experience the same force. Consider a positive test charge Q placed at point O just off the surface of the negative plate. In order to move it towards the positive plate we have to apply a force QE. The work done in moving the charge from the negative to the positive plate is,

$$\begin{aligned} W &= Fs \\ &= QEd, \end{aligned}$$

but from the definition of electrical potential,

$$W = VQ.$$

Equating these two expressions for the work done,

$$QEd = VQ,$$

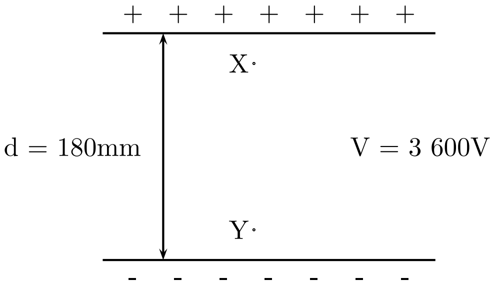
and so, rearranging,

$$E = \frac{V}{d}.$$

Worked Example 78

Parallel plates

Question:
Two charged parallel plates are at a distance of 180 mm from each other. The potential difference between them is 3600 V as shown in the diagram.



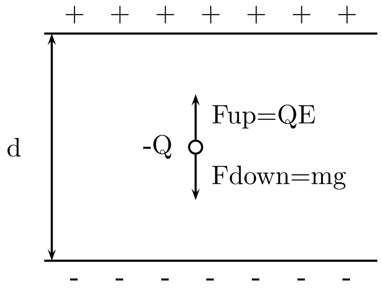


Figure 13.3: An oil drop suspended between oppositely charged parallel plates.

(a) If a small oil drop of negligible mass, carrying a charge of $+6.8 \times 10^{-9}$ C, is placed between the plates at point X, calculate the magnitude and direction of the electrostatic force exerted on the droplet.
(b) If the droplet is now moved to point Y, would the force exerted on it be bigger, smaller or the same as in (a)?
Answer:
(a)
Step 1 : (NOTE TO SELF: step is deprecated, use westep instead.)
First find the electric field strength between the plates,

$$\begin{aligned} E &= \frac{V}{d} \\ &= \frac{3600}{0.180} \\ &= 20000 \text{ N.C}^{-1} \text{ from the positive to the negative plate} \end{aligned}$$

Step 2 : (NOTE TO SELF: step is deprecated, use westep instead.)
Now the force exerted on the charge at X is,

$$\begin{aligned} F &= QE \\ &= (6.8 \times 10^{-9})(20000) \\ &= 1.36 \times 10^{-4} \text{ down} \end{aligned}$$

(b)
Step 3 : (NOTE TO SELF: step is deprecated, use westep instead.)
The same. Since the electric field strength is uniform, the force exerted on a charge is the same at all points between the plates.

13.5.3 Millikan's Oil-drop Experiment

Robert Millikan measured the charge on an electron by studying the motion of charged oil drops between oppositely charged parallel plates. Consider one such negative drop between the plates in Fig.13.3. Since this drop is negative, the electric field exerts an upward force on the drop. In addition to this upward force, gravity exerts a downward force on the drop. Millikan adjusted the electric field strength between the plates by varying the potential difference applied across the plates. In this way, Millikan was able to bring the drops to rest. At equlibrum,

$$\begin{aligned} F_{\text{up}} &= F_{\text{down}} \\ QE &= mg \end{aligned}$$

Since $E = \frac{V}{d}$,

$$Q\frac{V}{d} = mg,$$

and, therefore,

$$Q = \frac{mgd}{V}$$

Millikan found that all drops had charges which were multiples of 1.6×10^{-19} C. Since objects become charged by gaining or losing electrons, the charge on an electron must be -1.6×10^{-19} C. The magnitude of the electron's charge is denoted by e,

$$e = 1.6 \times 10^{-19}\ \text{C}$$

Worked Example 79

Charge

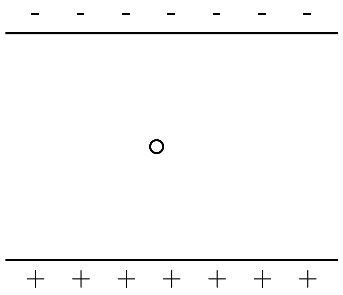
Question:
A metal sphere carries a charge of $+3.2 \times 10^{-8}$ C. How many electrons did it have to lose to attain its charge?
Answer:
Since the sphere is positive it lost electrons in the process of charging (when an object loses negative charges it is left positive). In fact, it lost,

$$\frac{3.2 \times 10^{-8}}{1.6 \times 10^{-19}} = 2 \times 10^{11}\ \text{electrons}$$

Worked Example 80

Millikan oil-drop experiment

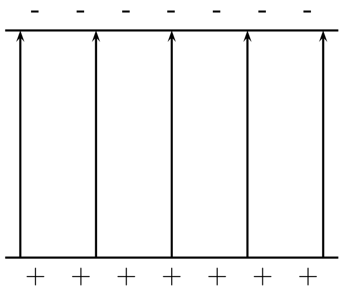
Question:
In a Millikan-type experiment a positively charged oil drop is placed between two horizontal plates, 20 mm apart, as shown.



The potential difference across the plates is 4000V. The drop has a mass of 1.2×10^{-14}kg and a charge of 8×10^{-19}C.
(a) Draw the electric field pattern between the two plates.
(b) Calculate:

1. the electric field intensity between the two plates.
2. the magnitude of the gravitational force acting on the drop.
3. the magnitude of the Coulomb force acting on the drop.

(c) The drop is observed through a microscope. What will the drop be seen to do? Explain.
(d) Without any further calculations, give two methods that could be used to make the drop remain in a fixed position.
Answer:
(a)



(b) 1.

$$\begin{aligned} E &= \frac{V}{d} \\ &= \frac{4000}{0.02} \\ &= 2 \times 10^5 \text{ V.m}^{-1} \text{ up} \end{aligned}$$

2.

$$\begin{aligned} F_{\text{grav}} &= mg \\ &= (1.2 \times 10^{-14})(10) \\ &= 1.2 \times 10^{-13}\text{N} \end{aligned}$$

3.

$$\begin{aligned} F_{\text{Coulomb}} &= QE \\ &= (8 \times 10^{-19})(s \times 10^{5}) \\ &= 1.6 \times 10^{-13}\text{N} \end{aligned}$$

(c) Since $F_{up} > F_{down}$, the drop accelerates upwards.
(d) The Coulomb force can be decreased by decreasing the electric field strength between the plates. Since $E = \frac{V}{d}$, this can be done either by increasing d or decreasing V.

13.6 Important Equations and Quantities

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
charge	q **or** Q	C (Coulomb)	$A.s$ **or** —	—
force	$\overrightarrow{F}$	N (Newton)	$\frac{kg.m}{s^2}$ **or** $kg.m.s^{-2}$	✓
mass	m	—	kg **or** —	—
acceleration	$\overrightarrow{a}$	—	$\frac{m}{s^2}$ **or** $m.s^{-2}$	✓
radial distance	r	—	m **or** —	—
electric field	$\overrightarrow{E}$	N/C **or** V/m	$\frac{kg.m}{A.s^3}$ **or** $kg.m.A^{-1}.s^{-3}$	✓
work	W	J	$\frac{kg.m^2}{s^2}$ **or** $kg.m^2.s^{-2}$	—
potential difference	V	V (Volt)	$\frac{kg.m^2}{A.s^3}$ **or** $kg.m^2.A^{-1}.s^{-3}$	—

Table 13.1: Units used in **Electrostatics**

Chapter 14

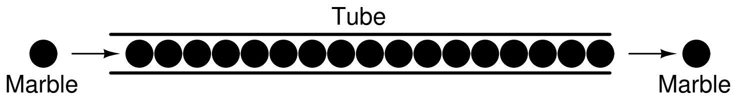
Electricity

Warning: We believe in experimenting and learning
about physics at every opportunity, BUT playing with
electricity can be **EXTREMELY DANGEROUS!**
Do not try to build home made circuits without someone who knows
if what you are doing is safe. Normal electrical outlets
are dangerous. Treat electricity with respect in your everyday life.
You will encounter electricity everyday for the rest of your life and to make sure you are able to

14.1 Flow of Charge

The normal motion of "free" electrons in a conductor has no particular direction or speed. However, electrons can be influenced to move in a coordinated fashion through a conductive material. This motion of electrons is what we call *electricity*, or *electric current*. This is in contrast to *static* electricity, which is an unmoving accumulation of electric charge. Just like water flowing through the emptiness of a pipe, electrons are able to move between the atoms of a conductor. The conductor may appear to be solid to our eyes, but any material composed of atoms is mostly empty space! The liquid-flow analogy is so fitting that the motion of electrons through a conductor is often referred to as a "flow."

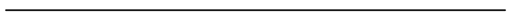
As each electron moves through a conductor, it pushes on the one ahead of it. This push of one electron on another makes all of the electrons move together as a group. The motion of the each electron in a conductor may be very slow. However, the starting and stopping of electron flow through a conductor is virtually instantaneous from one end of it to the other. As an analogy consider a tube filled end-to-end with marbles:



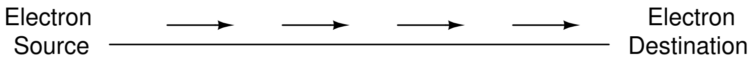
The tube is full of marbles, just as a conductor is full of free electrons. If a single marble is suddenly inserted into this full tube on the left-hand side, another marble will immediately try to exit the tube on the right. Even though each marble only traveled a short distance, the transfer of motion through the tube is virtually instantaneous from the left end to the right end. The nearly instantaneous transfer of motion through the tube occurs no matter how long the tube is. With electricity, the overall effect from one end of a conductor to the other is effectively instantaneous. Each individual electron, though, travels through the conductor at a *much* slower pace.

If we want electrons to flow in a certain direction to a certain place, we must provide the proper path for them to move. A path for electrons must be provided just as a plumber must install piping to get water to flow where he or she wants it to flow. *Wires* made of highly conductive metals such as copper or aluminum are used to form this path.

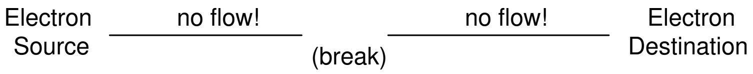
This means that there can be electric current *only* where there exists a continuous path of conductive material (wire) providing a path for electrons. In the marble analogy, marbles can flow into the left-hand side of the tube only if the tube is open on the right-hand side for marbles to flow out. If the tube is blocked on the right-hand side, the marbles will just "pile up" inside the tube. Marble "flow" will not occur if the tube is blocked. The same holds true for electric current: the continuous flow of electrons requires there be an unbroken path. Let's look at a diagram to illustrate how this works:



A thin, solid line (as shown above) is the conventional symbol for a continuous piece of wire. The wire is made of a conductive material, such as copper or aluminum. The wire's constituent atoms have many free electrons which can easily move through the wire. However, there will never be a continuous flow of electrons within this wire unless they have a place to come from and a place to go. Let's add an hypothetical electron "Source" and "Destination:"

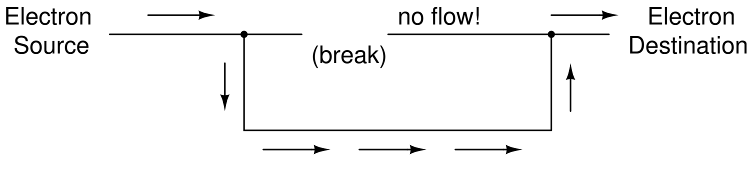


Now, with the Electron Source pushing new electrons into the wire on the left-hand side, electron flow through the wire can occur (as indicated by the arrows pointing from left to right). However, the flow will be interrupted if the conductive path formed by the wire is broken:



Air is an insulator that impedes the flow of electrons. An air gap separates the two pieces of wire, the path has now been broken, and electrons cannot flow from Source to Destination. This is like cutting a water pipe in two and capping off the broken ends of the pipe: water can't flow if there's no exit out of the pipe.

If we were to take another piece of wire leading to the Destination and connect it with the wire leading to the Source, we would once again have a continuous path for electrons to flow. The two dots in the diagram indicate physical (metal-to-metal) contact between the wire pieces:



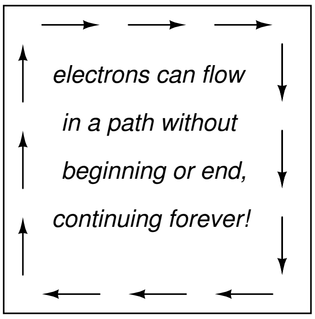
Now, we have continuity from the Source, to the newly-made connection, down, to the right, and up to the Destination. Please take note that the broken segment of wire on the right hand side has no electrons flowing through it. This is because it is no longer part of a complete path from Source to Destination.

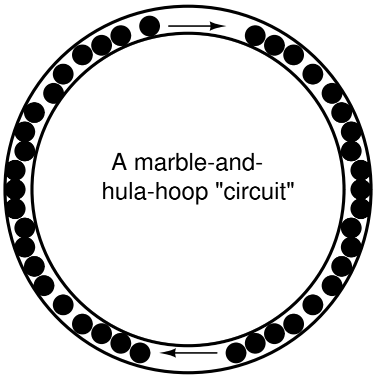
It is interesting to note that no ”wear” occurs within wires due to this electric current. This is in contrast to water-carrying pipes which are eventually corroded and worn by prolonged flows. Electrons do encounter some degree of friction as they move and this friction can generate heat in a conductor. This is a topic we'll discuss later.

14.2 Circuits

In order for the Source-and-Destination scheme to work, both would have to have a huge reservoir of electrons in order to sustain a continuous flow! Using the marble-and-tube analogy, the source and destination buckets would have to be large reservoirs to contain enough marble capacity for a ”flow” of marbles to be sustained.

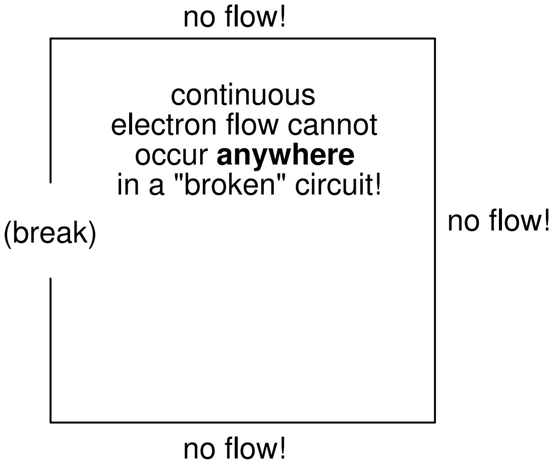
The answer to this paradox is found in the concept of a *circuit*: a never-ending looped pathway for electrons. If we take a wire, and loop it around so that it forms a continuous pathway, we have the means to support a uniform flow of electrons without having to resort to huge reservoirs.



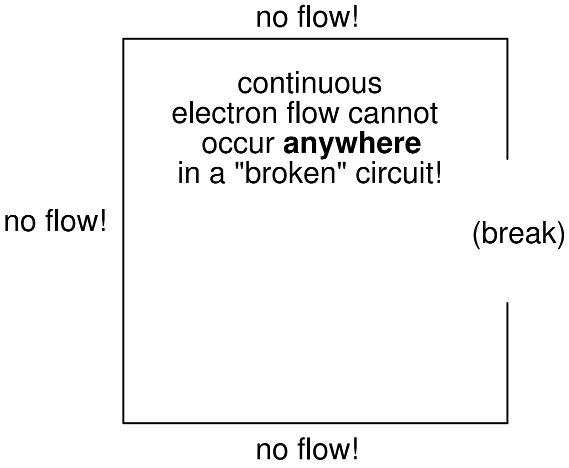


Each electron advancing clockwise in this circuit pushes on the one in front of it, which pushes on the one in front of it, and so on. A circuit is just like a hula-hoop filled with marbles. Now, we have the capability of supporting a continuous flow of electrons indefinitely without the need for reservoirs. All we need to maintain this flow is a continuous means of motivation for those electrons. This topic will be addressed in the next section of this chapter.

It must be realized that continuity is just as important in a circuit as it is in a straight piece of wire. Just as in the example with the straight piece of wire between the electron Source and Destination, any break in this circuit will prevent electrons from flowing through it:



An important principle to realize here is that *it doesn't matter where the break occurs.* Any discontinuity in the circuit will prevent electron flow throughout the entire circuit.



A combination of batteries and conductors with other components is called an electric circuit or circuit. The word circuit implies that you return to your starting point and this is an important property of electric circuits. They must contain a closed loop before charge can flow.

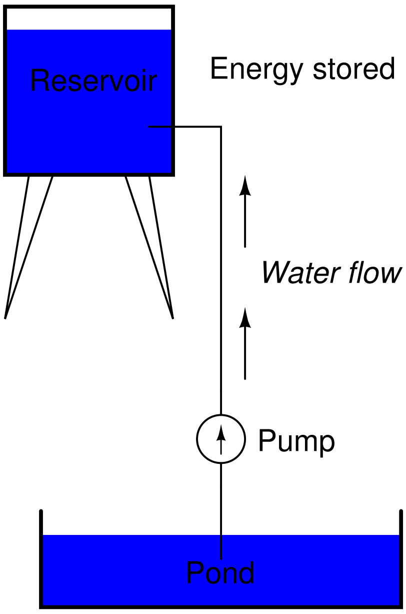
The simplest possible circuit is a battery with a single conductor. Now how do we form a closed loop with these two components? The battery has two terminals (connection points). One is called the positive terminal and one the negative. When we describe charge flowing we consider charges moving from the positive terminal of the battery around the conductor and back into the battery at the negative terminal. As much charge flows out of the positive terminal as flows into the negative terminal. Because of this, there is no build up of charge in the battery. The battery does work on the charges causing them to move round the circuit.

We have covered the topics of batteries and circuits but we need to draw these things to help keep the ideas clear in our minds. To do this we need to agree on how to draw things so that other people can understand what we are doing. We need a convention.

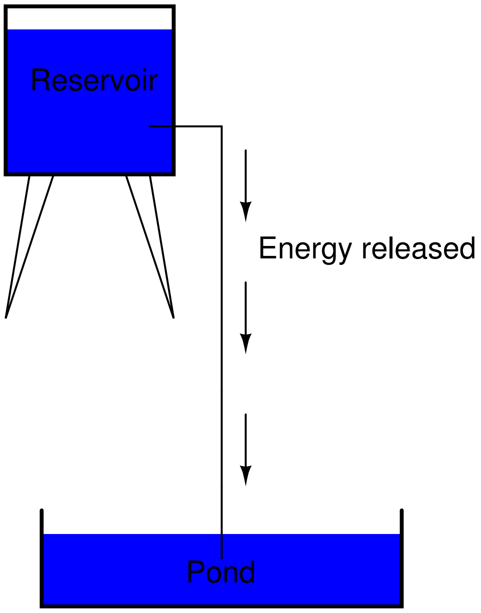
14.3 Voltage and current

As was previously mentioned, we need more than just a continuous path (circuit) before a continuous flow of electrons will occur. We also need some means to push these electrons around the circuit. Just like marbles in a tube or water in a pipe, it takes some kind of influencing force to initiate flow. With electrons, this force is the same force at work in static electricity: the force produced by an imbalance of electric charge.

The electric charge difference serves to store a certain amount of energy. This energy is not unlike the energy stored in a high reservoir of water that has been pumped from a lower-level pond:

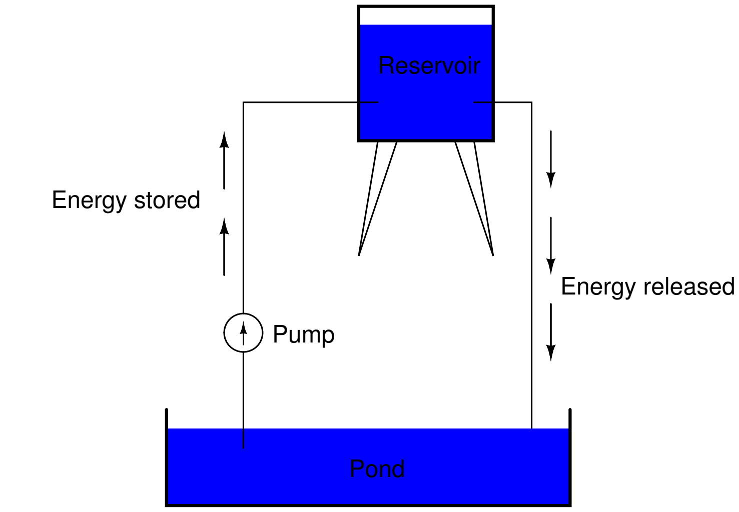


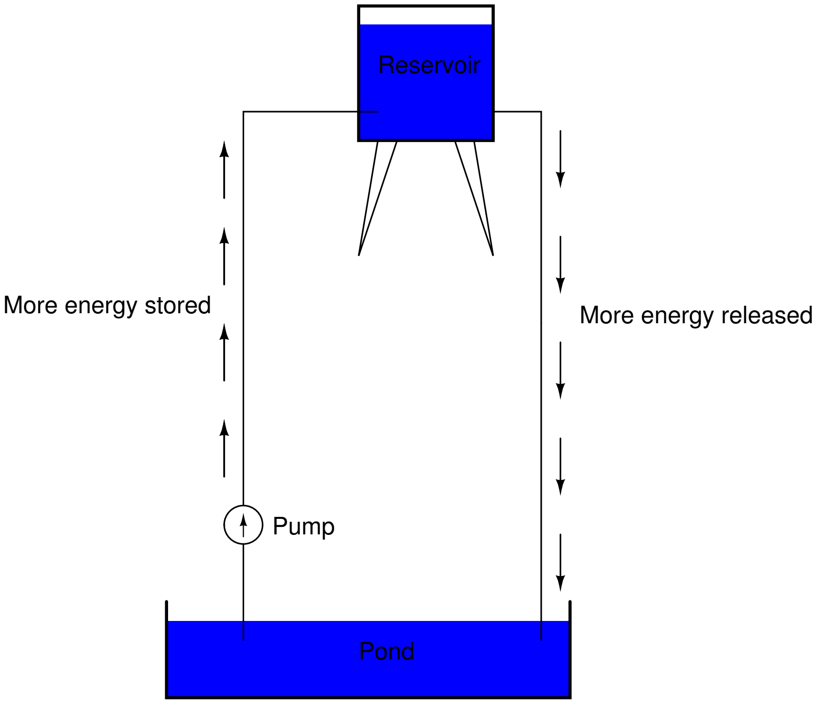
The influence of gravity on the water in the reservoir creates a force that attempts to move the water down to the lower level again. If a suitable pipe is run from the reservoir back to the pond, water will flow under the influence of gravity down from the reservoir, through the pipe:



It takes energy to pump that water from the low-level pond to the high-level reservoir. The movement of water through the piping back down to its original level constitutes a releasing of energy stored from previous pumping.

If the water is pumped to an even higher level, it will take even more energy to do so, and so more energy will be stored. The more energy is stored the more energy is released if the water is allowed to flow through a pipe back down again:





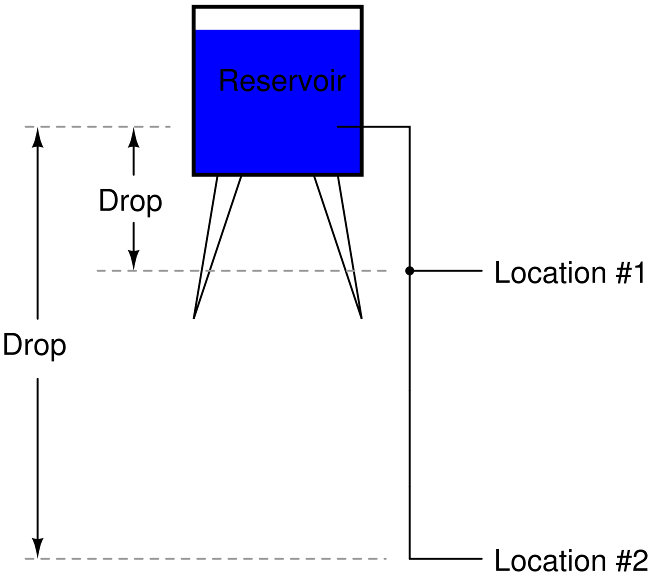
Electrons are not much different. If we "pump" electrons away from their normal "levels," we create a condition where a force exists as the electrons seek to re-establish their former positions. The force attracting electrons back to their original positions is analogous to the force gravity exerts on water in the reservoir. Just as gravity tries to draw water down to its former level, the force exerted on the electrons attracts them back to their former posisions.

Just as the pumping of water to a higher level results in energy being stored, "pumping" electrons to create an electric charge imbalance results in a certain amount of energy being stored in that imbalance. Providing a way for water to flow back down from the heights of the reservoir results in a release of that stored energy. Similarly, providing a way for electrons to flow back to their original "levels" results in a release of stored energy.

When the electrons are poised in that static condition (just like water sitting still, high in a reservoir), the energy stored there is called *potential energy*. It is given that name because it has the possibility (potential) of release that has not been fully realized yet.

This potential energy, stored in the form of an electric charge imbalance, can be expressed as a term called *voltage*. Technically, voltage is a measure of potential energy per unit charge of electrons, or something a physicist would call *specific potential energy*. Defined in the context of static electricity, voltage is the measure of work required to move a unit charge from one location to another. This work is against the force which tries to keep electric charges balanced. In the context of electrical power sources, voltage is the amount of potential energy available (work to be done) per unit charge, to move electrons through a conductor.

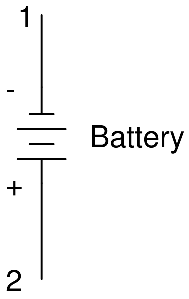
Voltage is an expression of potential energy. As such, it represents the possibility or potential for energy release as the electrons move from one "level" to another. Because of this, voltage is always referenced between two points. Consider the water reservoir analogy:



Because of the difference in the height of the drop, there's potential for much more energy to be released from the reservoir through the piping to location 2 than to location 1. The principle can be intuitively understood in dropping a rock: which results in a more violent impact, a rock dropped from a height of one foot, or the same rock dropped from a height of one kilometre? Obviously, the drop of greater height results in greater energy released (a more violent impact). We cannot assess the amount of stored energy in a water reservoir simply by measuring the volume of water. Similarly, we cannot predict the severity of a falling rock's impact simply from knowing the weight of the rock: in both cases we must also consider how *far* these masses will drop from their initial height. The amount of energy released by allowing a mass to drop is relative to the distance *between* its starting and ending points. Likewise, the potential energy available for moving electrons from one point to another is relative to those two points. Therefore, voltage is always expressed as a quantity *between* two points. Interestingly enough, the analogy of a mass potentially "dropping" from one height to another is such an apt model that voltage between two points is sometimes called a *voltage drop*.

Voltage can be generated in many ways. Chemical reactions, radiant energy, and the influence of magnetism on conductors are a few ways in which voltage may be produced. Respective examples of these three sources of voltage are batteries, solar cells, and generators. For now, we won't go into detail as to how each of these voltage sources works. The important thing is that we understand how voltage sources can be applied to create electron flow in a circuit.

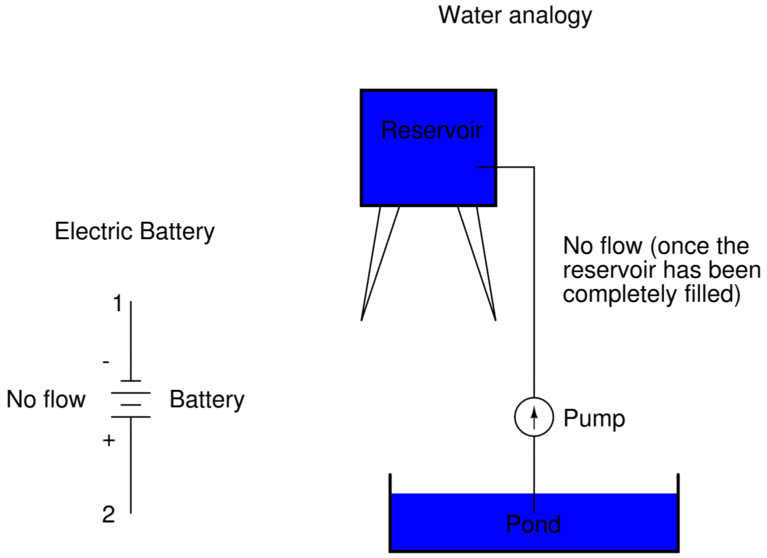
Let's take the symbol for a chemical battery and build a circuit step by step:



Any source of voltage, including batteries, have two points for electrical contact. In this case, we have point 1 and point 2 in the above diagram. The horizontal lines of varying length indicate that this is a battery. The horizontal lines further indicate the direction which this battery's voltage will try to push electrons through a circuit. The horizontal lines in the battery symbol appear separated, and so make it appear as if the battery is unable to serve as a path for electrons to move. This is no cause for concern: in real life, those horizontal lines represent metallic plates immersed in a liquid or semi-solid material that not only conducts electrons, but also generates the voltage to push them along by interacting with the plates.

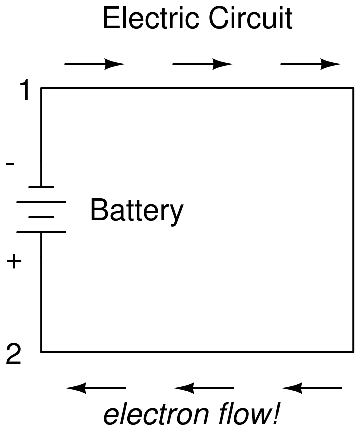
Notice the little "+" and "-" signs to the immediate left of the battery symbol. The negative (-) end of the battery is always the end with the shortest dash, and the positive (+) end of the battery is always the end with the longest dash. By convention, electrons are said to be "negatively" charged, so the negative end of a battery is that end which tries to push electrons out of it. Likewise, the positive end is that end which tries to attract electrons.

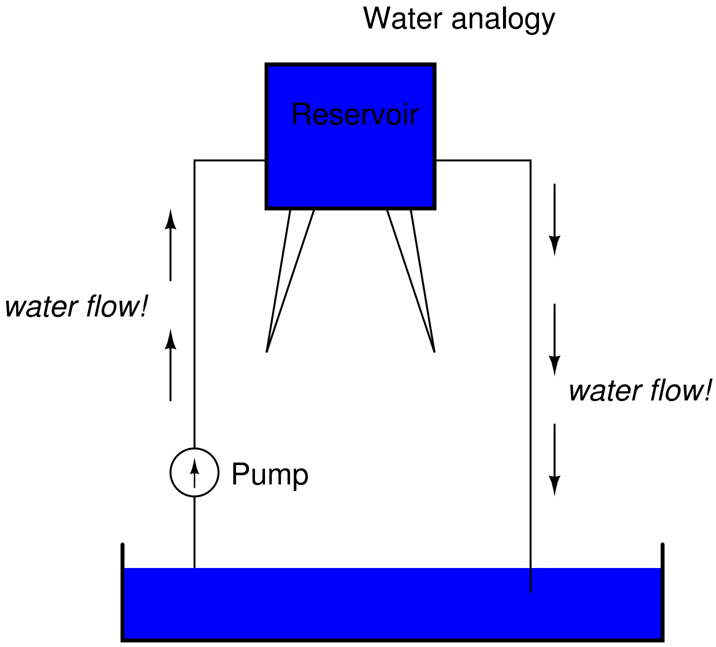
With the "+" and "-" ends of the battery not connected to anything, there will be voltage between those two points. However, there will be no flow of electrons through the battery, because there is no continuous path for the electrons to move.



The same principle holds true for the water reservoir and pump analogy: without a return pipe back to the pond, stored energy in the reservoir cannot be released in the form of water flow. Once the reservoir is completely filled up, no flow can occur, no matter how much pressure the pump may generate. There needs to be a complete path (circuit) for water to flow from the pond, to the reservoir, and back to the pond in order for continuous flow to occur.

We can provide such a path for the battery by connecting a piece of wire from one end of the battery to the other. Forming a circuit with a loop of wire, we will initiate a continuous flow of electrons in a clockwise direction:

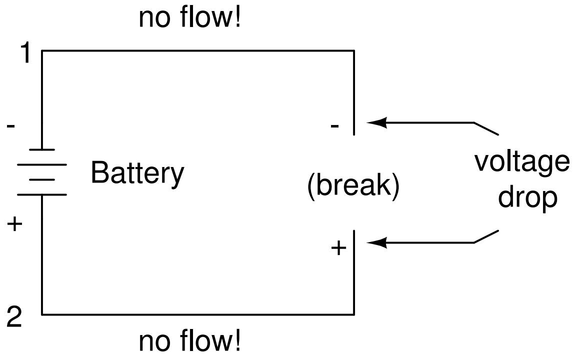




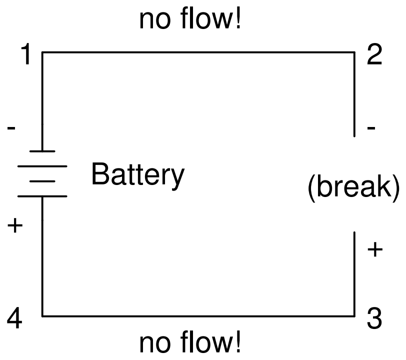
So long as the battery continues to produce voltage and the continuity of the electrical path isn't broken, electrons will continue to flow in the circuit. Following the metaphor of water moving through a pipe, this continuous, uniform flow of electrons through the circuit is called a *current*. So long as the voltage source keeps "pushing" in the same direction, the electron flow will continue to move in the same direction in the circuit. This single-direction flow of electrons is called a *Direct Current*, or DC.

Because electric current is composed of individual electrons flowing in unison through a conductor, just like marbles through a tube or water through a pipe, the amount of flow throughout a single circuit will be the same at any point. If we were to monitor a cross-section of the wire in a single circuit, counting the electrons flowing by, we would notice the exact same quantity per unit of time as in any other part of the circuit. The same quantity would be observed regardless of conductor length or conductor diameter.

If we break the circuit's continuity *at any point*, the electric current will cease in the entire loop. Futhermore, the full voltage produced by the battery will be manifested across the break, between the wire ends that used to be connected:

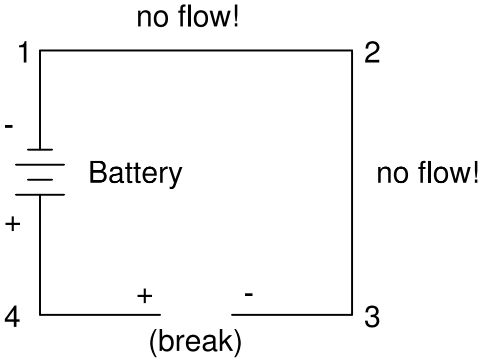


Notice the "+" and "-" signs drawn at the ends of the break in the circuit, and how they correspond to the "+" and "-" signs next to the battery's terminals. These markers indicate the direction that the voltage attempts to push electron flow. Remember that voltage is always relative between two points. Whether a point in a circuit gets labeled with a "+" or a "-" depends on the other point to which it is referenced. Take a look at the following circuit, where each corner of the loop is marked with a number for reference:



With the circuit's continuity broken between points 2 and 3, voltage dropped between points 2 and 3 is "-" for point 2 and "+" for point 3.

Now let's see what happens if we connect points 2 and 3 back together again, but place a break in the circuit between points 3 and 4:

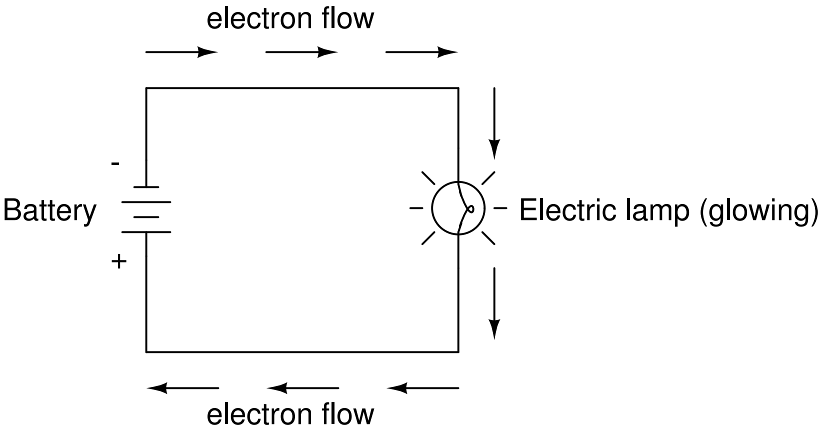


With the break between 3 and 4, the voltage drop between those two points is "+" for 4 and "-" for 3. Take special note of the fact that point 3's "sign" is opposite of that in the first example, where the break was between points 2 and 3 (where point 3 was labeled "+"). It is impossible for us to say that point 3 in this circuit will always be either "+" or "-", because the sign is not specific to a single point, but is always relative between two points!

14.4 Resistance

The circuit in the previous section is not a very practical one. In fact, it can be quite dangerous to directly connect the poles of a voltage source together with a single piece of wire. This is because the magnitude of electric current may be very large in such a *short circuit*, and the release of energy very dramatic (usually in the form of heat). Usually, electric circuits are constructed in such a way as to make practical use of that released energy, in as safe a manner as possible.

One practical and popular use of electric current is for the operation of electric lighting. The simplest form of electric lamp is a tiny metal "filament" inside of a clear glass bulb, which glows white-hot ("incandesces") with heat energy when sufficient electric current passes through it. Like the battery, it has two conductive connection points, one for electrons to enter and the other for electrons to exit. Connected to a source of voltage, an electric lamp circuit looks something like this:

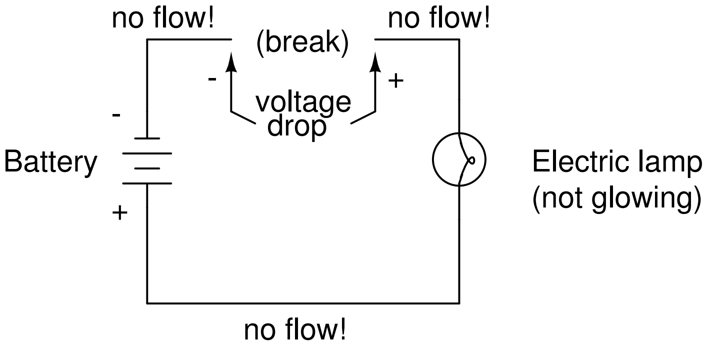


As the electrons work their way through the thin metal filament of the lamp, they encounter more opposition to motion than they typically would in a thick piece of wire. This opposition to electric current depends on the type of material, its cross-sectional area, and its temperature. It is technically known as *resistance*. It serves to limit the amount of current through the circuit with a given amount of voltage supplied by the battery. The "short circuit" where we had nothing but a wire joining one end of the voltage source (battery) to the other had not such a limiting resistance.

Interesting Fact: Materials known as conductors have a low resistance, while insulators have a very high one.

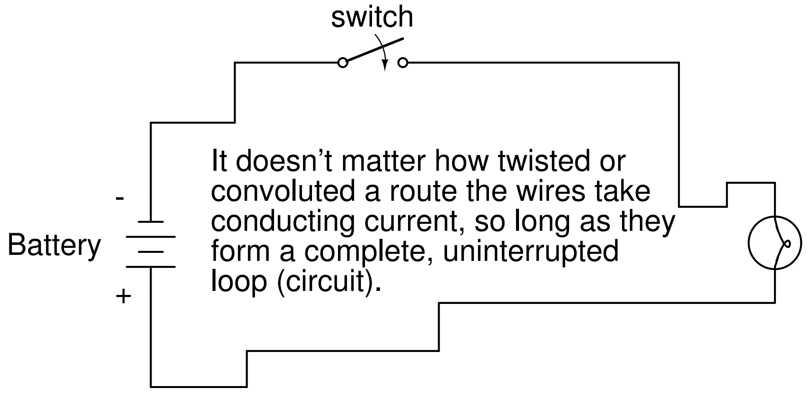
When electrons move against the opposition of resistance, "friction" is generated. Just like mechanical friction, the friction produced by electrons flowing against a resistance manifests itself in the form of heat. The concentrated resistance of a lamp's filament results in a relatively large amount of heat energy dissipated at that filament. This heat energy is enough to cause the filament to glow white-hot, producing light. The wires connecting the lamp to the battery hardly even get warm while conducting the same amount of current. This is because of their much lower resistance due to their larger cross-section.

As in the case of the short circuit, if the continuity of the circuit is broken at any point, electron flow stops throughout the entire circuit. With a lamp in place, this means that it will stop glowing:



As before, with no flow of electrons the entire potential (voltage) of the battery is available across the break, waiting for the opportunity of a connection to bridge across that break and permit electron flow again. This condition is known as an *open circuit*, where a break in the continuity of the circuit prevents current throughout. All it takes is a single break in continuity to "open" a circuit. Once any breaks have been connected once again and the continuity of the circuit re-established, it is known as a *closed circuit*.

What we see here is the basis for switching lamps on and off by switches. Because any break in a circuit's continuity results in current stopping throughout the entire circuit, we can use a device designed to intentionally break that continuity (called a *switch*). This switch can be mounted at any convenient location that we can run wires to. It controls the flow of electrons in the hole circuit:

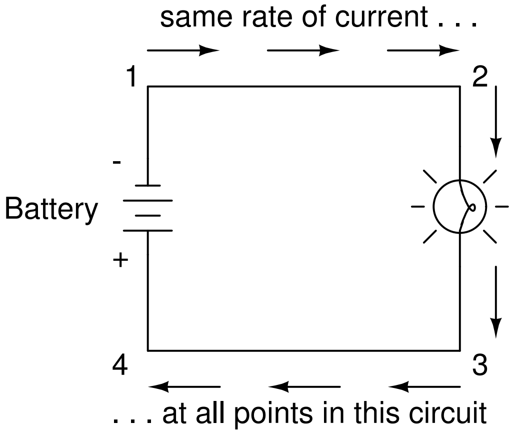

switch
It doesn't matter how twisted or convoluted a route the wires take conducting current, so long as they form a complete, uninterrupted loop (circuit).
-
Battery
+

This is how a switch mounted on the wall of a house can control a lamp that is mounted down a long hallway, or even in another room, far away from the switch. The switch itself is constructed of a pair of conductive contacts (usually made of some kind of metal) forced together by a mechanical lever actuator or pushbutton. When the contacts touch each other, electrons are able to flow from one to the other and the circuit's continuity is established; when the contacts are separated, electron flow from one to the other is prevented by the insulation of the air between, and the circuit's continuity is broken.

In keeping with the "open" and "closed" terminology of circuits, a switch that is making contact from one connection terminal to the other provides continuity for electrons to flow through, and is called a *closed* switch. Conversely, a switch that is breaking continuity won't allow electrons to pass through and is called an *open* switch. This terminology is often confusing to the new student of electronics, because the words "open" and "closed" are commonly understood in the context of a door, where "open" is equated with free passage and "closed" with blockage. With electrical switches, these terms have opposite meaning: "open" means no flow while "closed" means free passage of electrons.

14.5 Voltage and current in a practical circuit

Because it takes energy to force electrons to flow against the opposition of a resistance, there will be voltage manifested (or "dropped") between any points in a circuit with resistance between them. It is important to note that although the amount of current is uniform in a simple circuit, the amount of voltage between different sets of points in a single circuit may vary considerably:

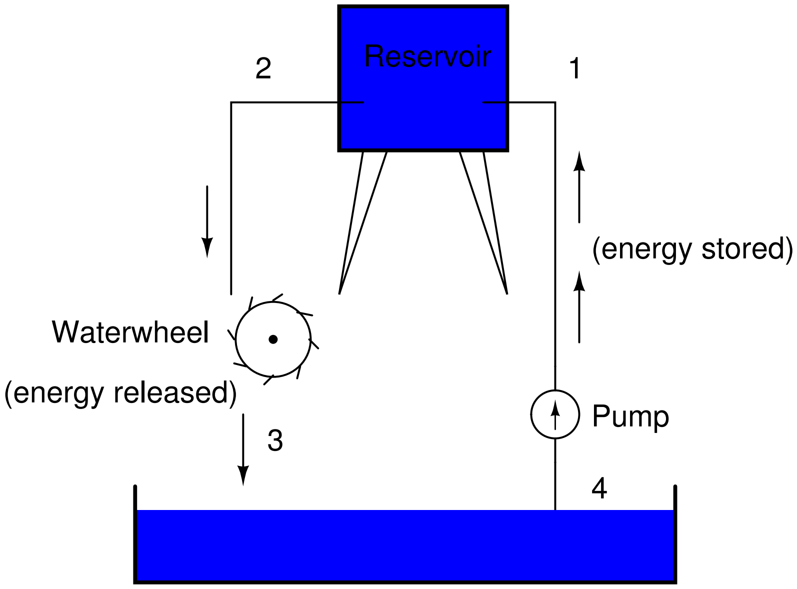

same rate of current . . .
1
2
-
Battery
+
4
3
. . . at all points in this circuit

(NOTE TO SELF: How do you actually measure currents and voltages? In the next two paragraphs there's a lot of ado what we 'see' at different points of a circuit, but I don't see anything!)

Take this circuit as an example. We labelled four points with the numbers 1, 2, 3, and 4. The amount of current conducted through the wire between points 1 and 2 is exactly the same as the amount of current conducted through the lamp (between points 2 and 3). This same quantity of current passes through the wire between points 3 and 4, and through the battery (between points 1 and 4).

However, we will find the voltage appearing between any two of these points to be directly proportional to the resistance within the conductive path between those two points. In a normal lamp circuit, the resistance of a lamp will be much greater than the resistance of the connecting wires. So we should expect to see a substantial amount of voltage drop between points 2 and 3, and only a very small one between points 1 and 2, or between 3 and 4. The voltage drop between points 1 and 4, of course, will be the full voltage offered by the battery. This will be only slightly higher than the voltage drop across the lamp (between points 2 and 3).

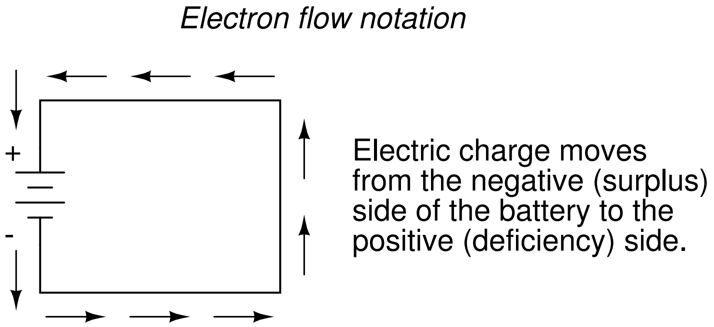
This, again, is analogous to the water reservoir system:



Between points 2 and 3, where the falling water is releasing energy at the water-wheel, there is a difference of pressure between the two points. This reflects the opposition to the flow of water through the water-wheel. From point 1 to point 2, or from point 3 to point 4, where water is flowing freely through reservoirs with little opposition, there is little or no difference of pressure (no potential energy). However, the rate of water flow in this continuous system is the same everywhere (assuming the water levels in both pond and reservoir are unchanging): through the pump, through the water-wheel, and through all the pipes. So it is with simple electric circuits: the rate of electron flow is the same at every point in the circuit, although voltages may differ between different sets of points.

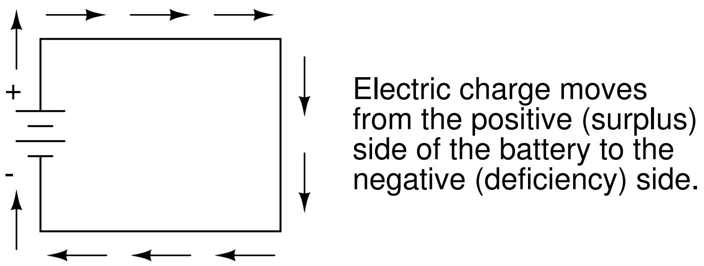
14.6 Direction of current flow in a circuit

We know now that the moving charges in an electrical ciruit are the negatively chargend electrons. These electrons naturally flow from the negative pole of a battery to the positive pole. This form of symbology became known as *electron flow* notation:


Electron flow notation
+
-
Electric charge moves
from the negative (surplus)
side of the battery to the
positive (deficiency) side.

However, for historical reasons the current flow in a circuit is conventionally denoted in the opposite direction. That is, it flows from the positive pole to the negative one. This became known as *conventional flow* notation:

Conventional flow notation


+
-
Electric charge moves from the positive (surplus) side of the battery to the negative (deficiency) side.

In conventional flow notation, we show the motion of charge according to the (technically incorrect) labels of + and −. This way the labels make sense, but the direction of charge flow is incorrect.

Does it matter, really, how we designate charge flow in a circuit? Not really, so long as we're consistent in the use of our symbols. You may follow an imagined direction of current (conventional flow) or the actual (electron flow) with equal success insofar as circuit analysis is concerned. Concepts of voltage, current, resistance, continuity, and even mathematical treatments such as Ohm's Law (section 2 (NOTE TO SELF: make this reference dynamic)) and Kirchhoff's Laws (section 6 (NOTE TO SELF: make this reference dynamic)) remain just as valid with either style of notation.

Aside: Benjamin Franklin made a conjecture regarding the direction of charge flow when rubbing smooth wax with rough wool. By assuming that the observed charges flow from the wax to the wool, he set the precedent for electrical notation that exists to this day. Because Franklin assumed electric charge moved in the opposite direction that it actually does, electrons are said to have a *negative* charge, and so objects he called "negative" (representing a deficiency of charge) actually have a surplus of electrons.

By the time the true direction of electron flow was discovered, the nomenclature of "positive" and "negative" had already been so well established in the scientific community that no effort was made to change it. It would have made more sense to call electrons "positive" in referring to "excess" charge. You see, the terms "positive" and "negative" are human inventions, and as such have no absolute meaning beyond our own conventions of language and scientific description. Franklin could have just as easily referred to a surplus of charge as "black" and a deficiency as "white", in which case scientists would speak of electrons having a "white" charge. However, because we tend to associate the word "positive" with "surplus" and "negative" with "deficiency," the standard label for electron charge does seem backward. As discussed above, many engineers decided to retain the old concept of electricity with "positive" referring to a surplus of charge, and label charge flow (current) accordingly.

14.7 How voltage, current, and resistance relate

First lets recap some of the ideas we have learnt so far. We will need these to understand how voltage, current and resistance relate.

An electric circuit is formed when a conductive path is created to allow free electrons to continuously move. This continuous movement of free electrons through the conductors of a circuit is called a *current*. It is often referred to in terms of "flow," just like the flow of a liquid through a hollow pipe. The force motivating electrons to "flow" in a circuit is called *voltage*. Voltage is a specific measure of potential energy that is always relative between two points. When we speak of a certain amount of voltage being present in a circuit, we are referring to the measurement of how much *potential* energy exists to move electrons from one particular point in that circuit to another particular point. Without reference to *two* particular points, the term "voltage" has no meaning.

Free electrons tend to move through conductors with some degree of friction, or opposition to motion. This opposition to motion is more properly called *resistance*. The amount of current in a circuit depends on the amount of voltage available to motivate the electrons (NOTE TO SELF: Motivated electrons?), and also the amount of resistance in the circuit to oppose electron flow. Just like voltage, resistance is a quantity relative between two points. For this reason, the quantities of voltage and resistance are often stated as being "between" or "across" two points in a circuit.

To be able to make meaningful statements about these quantities in circuits, we need to be able to describe their quantities in the same way that we might quantify mass, temperature, volume, length, or any other kind of physical quantity. For mass we might use the units of "pound" or "gram". For temperature we might use degrees Fahrenheit or degrees Celsius. Here are the standard units of measurement for electrical current, voltage, and resistance:

Quantity	Symbol	Unit of Measurement	Abbreviation of Unit
Current	I	Ampere	A
Voltage	V (or E)	volt	V
Resistance	R	Ohm	Ω
Charge	Q	coulomb	C

The "symbol" given for each quantity is the standard alphabetical letter used to represent that quantity in an algebraic equation. Standardized letters like these are common in the disciplines of physics and engineering, and are internationally recognized. The "unit abbreviation" for each quantity represents the alphabetical symbol used as a shorthand notation for its particular unit of measurement. And, yes, that strange-looking "horseshoe" symbol is the capital Greek letter Ω (called omega), just a character in a *foreign* alphabet (apologies to any Greek readers here).

Aside: Each unit of measurement is named after a famous experimenter in electricity: The *amp* after the Frenchman Andre M. Ampere, the *volt* after the Italian Alessandro Volta, and the *ohm* after the German Georg Simon Ohm.
The mathematical symbol for each quantity is meaningful as well. The "R" for resistance and the "V" for voltage are both self-explanatory. The "I" is thought to have been meant to represent "Intensity" (of electron flow). The other symbol for voltage, "E", stands for "Electromotive force". The symbols "E" and "V" are interchangeable for the most part, although some texts reserve "E" to represent voltage across a source (such as a battery or generator) and "V" to represent voltage across anything else.

One foundational unit of electrical measurement is the unit of the *coulomb*. It is a measure of electric charge proportional to the number of electrons in an imbalanced state. One coulomb of charge is rougly equal to the charge of 6,250,000,000,000,000,000 electrons. The symbol for electric charge quantity is the capital letter ”Q”, with the unit of coulombs abbreviated by the capital letter ”C”. It so happens that the unit for electron flow, the ampere, is equal to 1 coulomb of electrons passing by a given point in a circuit in 1 second of time. Cast in these terms, current is the *rate of electric charge motion* through a conductor.

As stated before, voltage is the measure of *potential energy per unit charge* available to motivate electrons from one point to another. Before we can precisely define what a ”volt” is, we must understand how to measure this quantity we call ”potential energy”. The general metric unit for energy of any kind is the *joule*, equal to the amount of work performed by a force of 1 newton exerted through a motion of 1 meter (in the same direction). (NOTE TO SELF: Make a reference to the Mechanics chapter.) Defined in these scientific terms, 1 volt is equal to 1 joule of electric potential energy per (divided by) 1 coulomb of charge. Thus, a 9 volt battery releases 9 joules of energy for every coulomb of electrons moved through a circuit.

These units and symbols for electrical quantities will become very important to know as we begin to explore the relationships between them in circuits. The first, and perhaps most important, relationship between current, voltage, and resistance is called Ohm's Law. It states that the amount of electric current through a metal conductor in a circuit is directly proportional to the voltage impressed across it, for any given temperature. It can be expressed in the form of a simple equation, describing how voltage, current, and resistance interrelate:

$$V = I \cdot R\,. \tag{14.1}$$

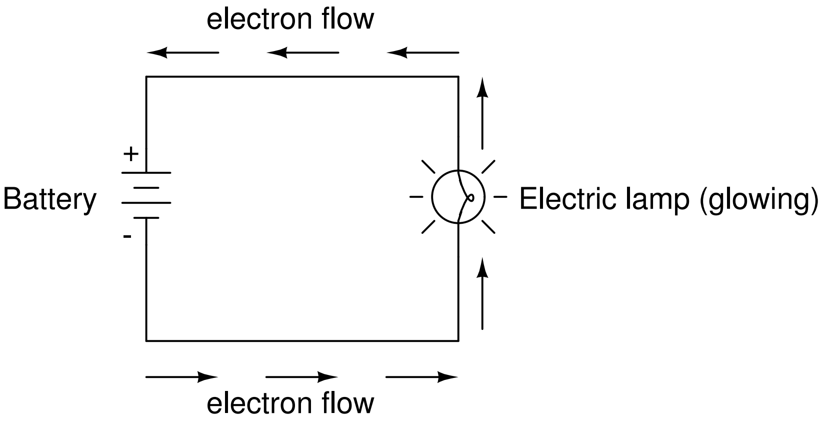
In this algebraic expression, voltage (V) is equal to current (I) multiplied by resistance (R).

Aside: Georg Simon Ohm published *his* law in his 1827 paper, *The Galvanic Circuit Investigated Mathematically.*

Using algebra techniques, we can manipulate this equation into two variations, solving for I and for R, respectively:

$$I = \frac{V}{R} \quad R = \frac{V}{I}\,. \tag{14.2}$$

Let's see how these equations might work to help us analyze simple circuits:

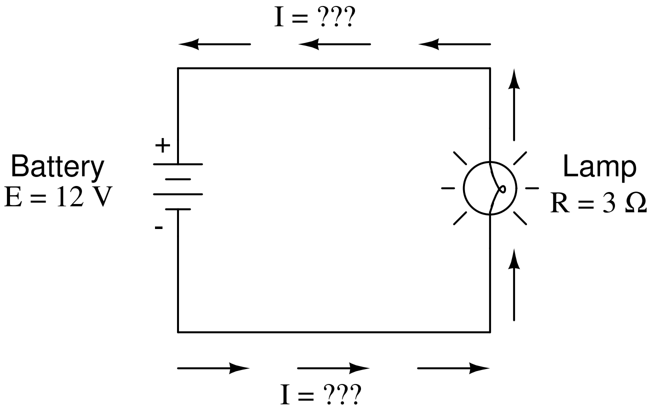


(NOTE TO SELF: replace E by V)

In the above circuit, there is only one source of voltage (the battery, on the left) and only one source of resistance to current (the lamp, on the right). This makes it very easy to apply Ohm's Law. If we know the values of any two of the three quantities (voltage, current, and resistance) in this circuit, we can use Ohm's Law to determine the third. In this first example, we will calculate the amount of current (I) in a circuit, given values of voltage (V) and resistance (R):

Worked Example 81

Question: What is the amount of current (I) in this circuit?



(NOTE TO SELF: replace E by V)

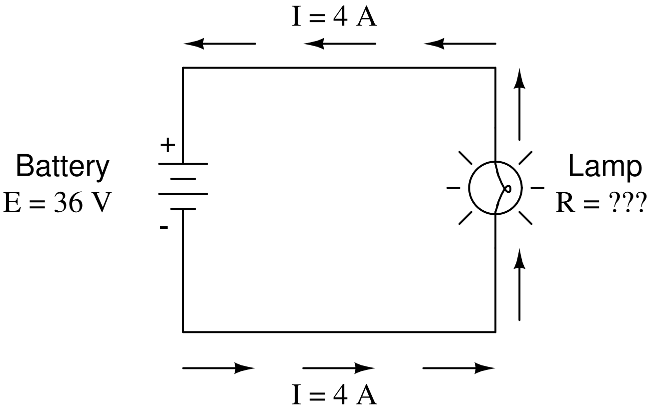
Answer:

$$I = \frac{V}{R} = \frac{12\,\mathrm{V}}{3\,\Omega} = 4\,\mathrm{A}\,. \tag{14.3}$$

In the second example, we will calculate the amount of resistance (R) in a circuit, given values of voltage (V) and current (I):

Worked Example 82

Question: What is the amount of resistance (R) offered by the lamp?



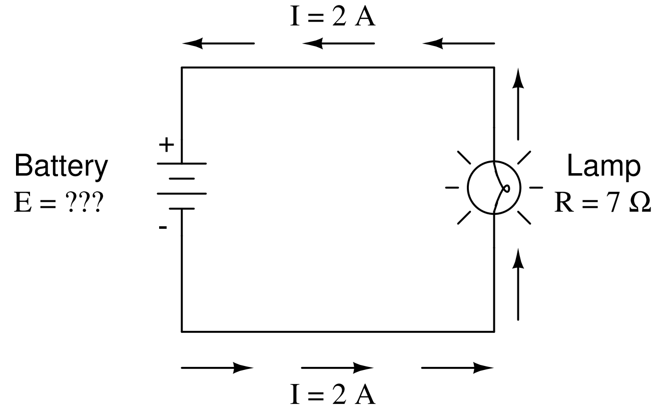
(NOTE TO SELF: replace E by V)
Answer:

$$R = \frac{V}{I} = \frac{36\,\mathrm{V}}{4\,\mathrm{A}} = 9\,\Omega\,. \tag{14.4}$$

In the last example, we will calculate the amount of voltage supplied by a battery, given values of current (I) and resistance (R):

Worked Example 83

Question: What is the amount of voltage provided by the battery?



(NOTE TO SELF: replace E by V)
Answer:

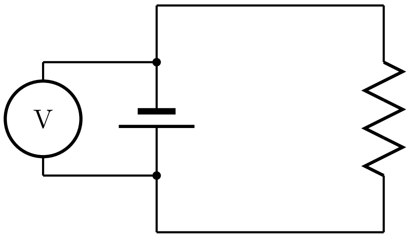
$$V = I \cdot R = (2\,\mathrm{A}) \cdot (7\,\Omega) = 14\,\mathrm{V}\,. \tag{14.5}$$

Ohm's Law is a very simple and useful tool for analyzing electric circuits. It is used so often in the study of electricity and electronics that it needs to be committed to memory by the serious student. All you need to do is commit $V = I \cdot R$ to memory and derive the other two formulae from that when you need them!

14.8 Voltmeters, ammeters, and ohmmeters

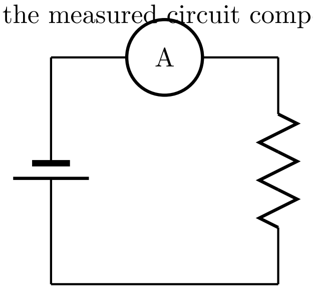
As we have seen in previous sections, an electric circuit is made up of a number of different components such as batteries and resistors. In electronics, there are many types of meters used to measure the properties of the individual components of an electric circuit. For example, one may be interested in measuring the amount of current flowing through a circuit, or measure the voltage provided by a battery. In this section we will discuss the practical usage of voltmeters, ammeters, and ohmmeters.

A voltmeter is an instrument for measuring the voltage between two points in an electric circuit. In analogy with a water circuit, a voltmeter is like a meter designed to measure pressure difference. Since one is interested in measuring the voltage between two points in a circuit, a voltmeter must be connected in *parallel* with the portion of the circuit on which the measurement is made:

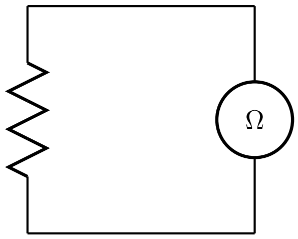


The above illustration shows a voltmeter connected in parallel with a battery. One lead of the voltmeter is connected to one end of the battery and the other lead is connected to the opposite end. The voltmeter may also be used to measure the voltage across a resistor or any other component of a circuit that has a voltage drop.

An ammeter is an instrument used to measure the flow of electric current in a circuit. Since one is interested in measuring the current flowing *through* a circuit component, the ammeter must be connected in *series* with the measured circuit component:



An ohmmeter is an instrument for measuring electrical resistance. The basic ohmmeter can function much like an ammeter. The ohmmeter works by suppling a constant voltage to the resistor and measuring the current flowing through it. The measured current is then converted into a corresponding resistance reading through Ohm's law. One cautionary detail needs to be mentioned with regard to ohmmeters: they only function correctly when measuring resistance that is not being powered by a voltage or current source. In other words, you cannot measure the resistance of a component that is already connected to a circuit. The reason for this is simple: the ohmmeter's accurate indication depends only on its own source of voltage. The presence of **any other** voltage across the measured circuit component interferes with the ohmmeter's operation. The circuit diagram below shows an ohmmeter solely connected with a resistor:



The table below summarizes the use of each measuring instrument that we discussed and the way it should be connected to a circuit component.

Instrument	Measured Quantity	Proper Connection
Voltmeter	Voltage	In Parallel
Ammeter	Current	In Series
Ohmmeter	Resistance	Only with Resistor

14.9 An analogy for Ohm's Law

In our water-and-pipe analogy, Ohm's Law also exists. Think of a water pump that exerts pressure (voltage) to push water around a "circuit" (current) through a restriction (resistance). If the resistance to water flow stays the same and the pump pressure increases, the flow rate must also increase.

$$\overset{\uparrow}{V}=\overset{\uparrow}{I}\ R$$

If the pressure stays the same and the resistance increases (making it more difficult for the water to flow), then the flow rate must decrease:

$$V=\overset{\downarrow}{I}\overset{\uparrow}{R}$$

If the flow rate stays the same while the resistance to flow decreases, the required pressure from the pump decreases:

$$\overset{\downarrow}{V}= I\ \overset{\downarrow}{R}$$

As odd as it may seem, the actual mathematical relationship between pressure, flow, and resistance is actually more complex for fluids like water than it is for electrons. If you pursue further studies in physics, you will discover this for yourself. Thankfully for the electronics student, the mathematics of Ohm's Law is very simple.

14.10 Power in electric circuits

In addition to voltage and current, there is another measure of free electron activity in a circuit: *power*. The concept of *power* was introduced in Chapter 8. Basically, it is a measure of how rapidly a standard amount of *work* is done.

In electric circuits, power is a function of both voltage and current:

$$P = IV.$$

So power (P) is exactly equal to current (I) multiplied by voltage (V) and there is no extra constant of proportionality. The unit of measurement for power is the *Watt* (abbreviated W).

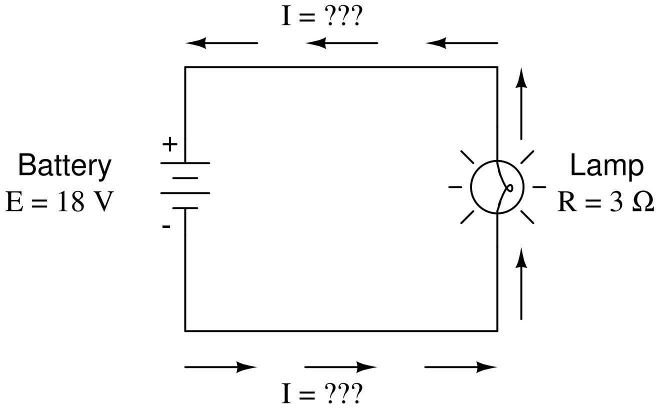
> *Aside:* You can verify for yourself that the eqution for power in an electric cicuit makes sense. Remember that voltage is the specific work (or potential energy) per unit charge, while current is the amount electric charge that flow though a conductor per time unit. So the product of those two qunatities is the oumount of work per time unit, which is exactly the power.

It is important to realise that only the combination of a voltage drop and the flow of current corresponds to power. So, a circuit with high voltage and low current may be dissipating the same amount of power as a circuit with low voltage and high current.

In an open circuit, where voltage is present between the terminals of the source and there is zero current, there is *zero* power dissipated, no matter how great that voltage may be. Since $P = IV$ and $I = 0$, the power dissipated in any open circuit must be zero.

14.11 Calculating electric power

We've seen the formula for determining the power in an electric circuit: by multiplying the voltage in volts by the current in Ampères we arrive at an answer in watts." Let's apply this to a circuit example:



In the above circuit, we know we have a battery voltage of 18 Volts and a lamp resistance of 3 Ω. Using Ohm's Law to determine current, we get:

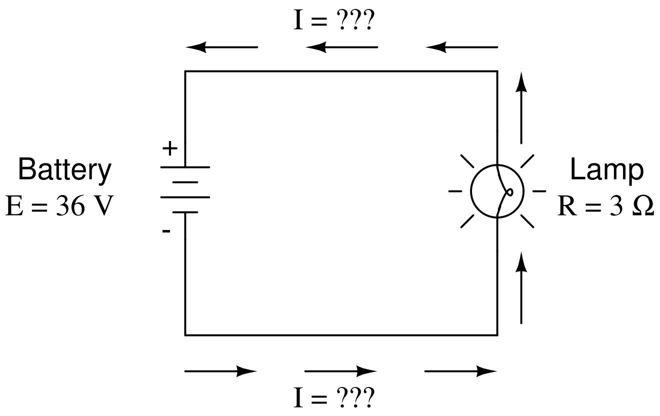
$$I = \frac{V}{R} = \frac{18\text{V}}{3\Omega} = 6\text{A}.$$

Now that we know the current, we can take that value and multiply it by the voltage to determine power:

$$P = IV = (6\text{A})(18\text{V} = 108\text{W}.$$

Answer: the lamp is dissipating (releasing) 108 W of power, most likely in the form of both light and heat.

Let's try taking that same circuit and increasing the battery voltage to see what happens:



Since the resistance stays the same, the current will increase when we increase the voltage:

$$I = \frac{V}{R} = \frac{36\text{V}}{3\Omega} = 12\text{A}.$$

Note that Ohm's Law is linear, so the current exactly doubles when we double the voltage. Now, let's calculate the power:

$$P = IV = (12\text{A})(36\text{V}) = 432\text{W}.$$

Notice that the power has increased just as we might have suspected, but it increased quite a bit more than the current. Why is this? Because power is a function of voltage multiplied by current, and *both* voltage and current doubled from their previous values, the power will increase by a factor of 2 x 2, or 4: the ratio of the new power 432 W and the old power 108 W, is exactly 4.

We could in fact have arrived at this result without the intermediate step of calculating the current. From

$$I = \tfrac{V}{R} \quad \text{and} \quad P = IV$$

we can expres power directly as a function of voltage:

$$P = \tfrac{V}{R}V = \tfrac{V^2}{R}.$$

The analogous relation between power and current is

$$P = IIR = I^2R.$$

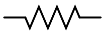
Interesting Fact: It was James Prescott Joule, not Georg Simon Ohm, who first discovered the mathematical relationship between power dissipation and current through a resistance. This discovery, published in 1841, followed the form of the last equation (P = I^2R), and is properly known as Joule's Law. However, these power equations are so commonly associated with the Ohm's Law equations relating voltage, current, and resistance ($V = IR$; $I = V/R$; and $R = V/I$) that they are frequently credited to Ohm.

14.12 Resistors

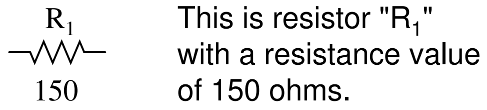
Because the relationship between voltage, current, and resistance in any circuit is so regular, we can reliably control any variable in a circuit simply by controlling the other two. Perhaps the easiest variable in any circuit to control is its resistance. This can be done by changing the material, size, and shape of its conductive components (remember how the thin metal filament of a lamp created more electrical resistance than a thick wire?).

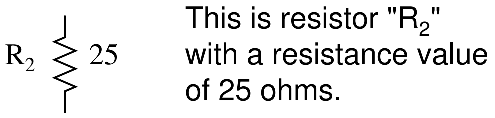
Special components called *resistors* are made for the express purpose of creating a precise quantity of resistance for insertion into a circuit. They are typically constructed of metal wire or carbon, and engineered to maintain a stable resistance value over a wide range of environmental conditions. Unlike lamps, they do not produce light, but they do produce heat as electric power is dissipated by them in a working circuit. Typically, though, the purpose of a resistor is not to produce usable heat, but simply to provide a precise quantity of electrical resistance.

The most common schematic symbol for a resistor is a zig-zag line:

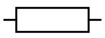


Resistor values in ohms are usually shown as an adjacent number, and if several resistors are present in a circuit, they will be labeled with a unique identifier number such as R_1, R_2, R_3, etc. As you can see, resistor symbols can be shown either horizontally or vertically:

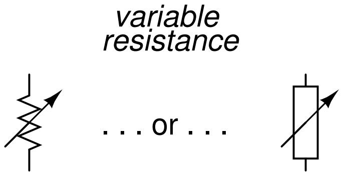




In keeping more with their physical appearance, an alternative schematic symbol for a resistor looks like a small, rectangular box:



Resistors can also be shown to have varying rather than fixed resistances. This might be for the purpose of describing an actual physical device designed for the purpose of providing an adjustable resistance, or it could be to show some component that just happens to have an unstable resistance:



In fact, any time you see a component symbol drawn with a diagonal arrow through it, that component has a variable rather than a fixed value. This symbol "modifier" (the diagonal arrow) is standard electronic symbol convention.

Aside: In practice, resistors are not only rated in terms of their resistance in ohms, but also in term the amount of power they can dissipate in watts. Resistors dissipate heat as the electric currents through them overcome the "friction" of their resistance and can in fact become quite hot in actual applications. Most resistors found in small electronic devices such as portable radios are rated at 1/4 (0.25) watt or less. The power rating of any resistor is roughly proportional to its physical size. Also note how resistances (in ohms) have nothing to do with size!

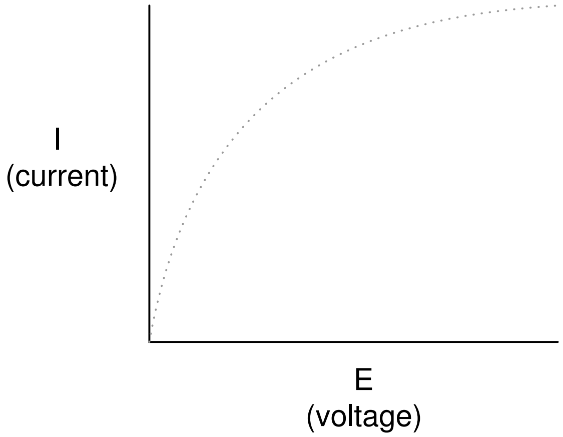
14.13 Nonlinear conduction

Ohm's Law is a powerful tool for analyzing electric circuits, but it has a practical limitation. In the application of Ohm's Law, we alwasy assume that the restistance does not change as a function of voltage and current. For most conductors, this is a reasonable approximation as long ads the temperature does not change too much.

In a normal lightbulb, the resistance of the filament wire will increase dramatically as it warms from room temperature to operating temperature. If we increase the supply voltage in a real lamp circuit, the resulting increase in current causes the filament to increase in temperature, which increases its resistance. This effectively limits the increase in current. Consequently, voltage and current do not follow the simple equation $I = V/R$, with a constant R (of 3 Ω ion our example). The lamp's filament resistance does not remain stable for different currents.

The phenomenon of resistance changing with variations in temperature is one shared by almost all metals, of which most wires are made. For most applications, these changes in resistance are small enough to be ignored. In the application of metal lamp filaments, which increase a lot in temperature (up to about 1000^{o}C, and starting from room temperature) the change is quite large.

A more realistic analysis of a lamp circuit over several different values of battery voltage would generate a plot of this shape:

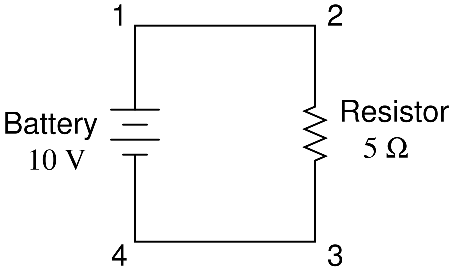


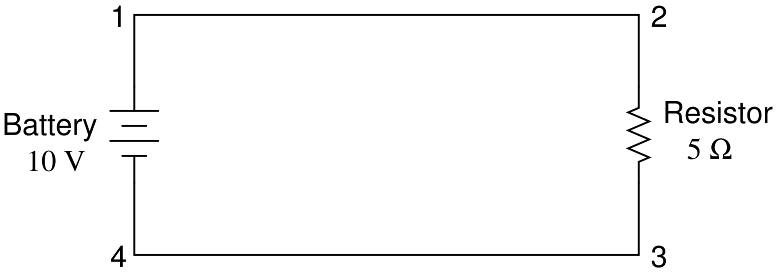
The plot is no longer a straight line. It rises sharply on the left, as voltage increases from zero to a low level. As it progresses to the right we see the line flattening out, the circuit requiring greater and greater increases in voltage to achieve equal increases in current.

If we apply Ohm's Law to find the resistance of this lamp circuit with the voltage and current values plotted above, the calculated values will change with voltage or curreny. We could say that the resistance here is *nonlinear*, increasing with increasing current and voltage. The nonlinearity is caused by the effects of high temperature on the metal wire of the lamp filament.

14.14 Circuit wiring

So far, we've been analyzing single-battery, single-resistor circuits with no regard for the connecting wires between the components, so long as a complete circuit is formed. Does the wire length or circuit "shape" matter to our calculations? Let's look at a couple of circuit configurations and find out:

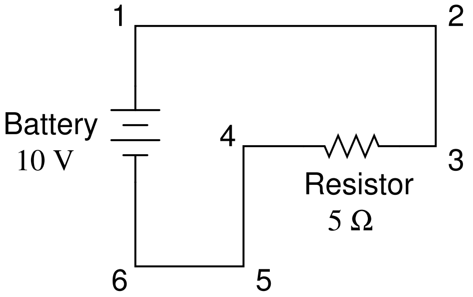




When we draw wires connecting points in a circuit, we usually assume those wires have negligible resistance. As such, they contribute no appreciable effect to the overall resistance of the circuit, and so the only resistance we have to contend with is the resistance in the components. In the above circuits, the only resistance comes from the 5 Ω resistors, so that is all we will consider in our calculations. In real life, metal wires actually *do* have resistance (and so do power sources!), but those resistances are generally so much smaller than the resistance present in the other circuit components that they can be safely ignored.

If connecting wire resistance is very little or none, we can regard the connected points in a circuit as being *electrically common*. That is, points 1 and 2 in the above circuits may be physically joined close together or far apart, and it doesn't matter for any voltage or resistance measurements relative to those points. The same goes for points 3 and 4. It is as if the ends of the resistor were attached directly across the terminals of the battery, so far as our Ohm's Law calculations and voltage measurements are concerned. This is useful to know, because it means you can re-draw a circuit diagram or re-wire a circuit, shortening or lengthening the wires as desired without appreciably impacting the circuit's function. All that matters is that the components attach to each other in the same sequence.

It also means that voltage measurements between sets of "electrically common" points will be the same. That is, the voltage between points 1 and 4 (directly across the battery) will be the same as the voltage between points 2 and 3 (directly across the resistor). Take a close look at the following circuit, and try to determine which points are common to each other:



Here, we only have 2 components excluding the wires: the battery and the resistor. Though the connecting wires take a convoluted path in forming a complete circuit, there are several electrically common points in the electrons' path. Points 1, 2, and 3 are all common to each other, because they're directly connected together by wire. The same goes for points 4, 5, and 6.

The voltage between points 1 and 6 is 10 volts, coming straight from the battery. However, since points 5 and 4 are common to 6, and points 2 and 3 common to 1, that same 10 volts also exists between these other pairs of points:

```
Between points 1 and 4 = 10 volts
Between points 2 and 4 = 10 volts
Between points 3 and 4 = 10 volts (directly across the resistor)
Between points 1 and 5 = 10 volts
Between points 2 and 5 = 10 volts
Between points 3 and 5 = 10 volts
Between points 1 and 6 = 10 volts (directly across the battery)
Between points 2 and 6 = 10 volts
Between points 3 and 6 = 10 volts
```

Since electrically common points are connected together by (zero resistance) wire, there is no significant voltage drop between them regardless of the amount of current conducted from one to the next through that connecting wire. Thus, if we were to read voltages between common points, we should show (practically) zero:

```
Between points 1 and 2 = 0 volts    Points 1, 2, and 3 are
Between points 2 and 3 = 0 volts     electrically common
Between points 1 and 3 = 0 volts
Between points 4 and 5 = 0 volts    Points 4, 5, and 6 are
Between points 5 and 6 = 0 volts     electrically common
Between points 4 and 6 = 0 volts
```

This makes sense mathematically, too. With a 10 volt battery and a 5 Ω resistor, the circuit current will be 2 amps. With wire resistance being zero, the voltage drop across any continuous stretch of wire can be determined through Ohm's Law as such:

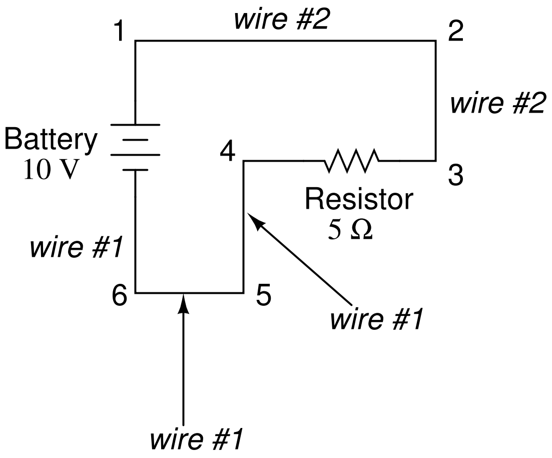
$$E = I R$$

$$E = (2\ A)(0\ \Omega)$$

$$E = 0\ V$$

It should be obvious that the calculated voltage drop across any uninterrupted length of wire in a circuit where wire is assumed to have zero resistance will always be zero, no matter what the magnitude of current, since zero multiplied by anything equals zero.

Because common points in a circuit will exhibit the same relative voltage and resistance measurements, wires connecting common points are often labeled with the same designation. This is not to say that the *terminal* connection points are labeled the same, just the connecting wires. Take this circuit as an example:



Points 1, 2, and 3 are all common to each other, so the wire connecting point 1 to 2 is labeled the same (wire 2) as the wire connecting point 2 to 3 (wire 2). In a real circuit, the wire stretching from point 1 to 2 may not even be the same color or size as the wire connecting point 2 to 3, but they should bear the exact same label. The same goes for the wires connecting points 6, 5, and 4.

Knowing that electrically common points have zero voltage drop between them is a valuable troubleshooting principle. If I measure for voltage between points in a circuit that are supposed to be common to each other, I should read zero. If, however, I read substantial voltage between those two points, then I know with certainty that they cannot be directly connected together.

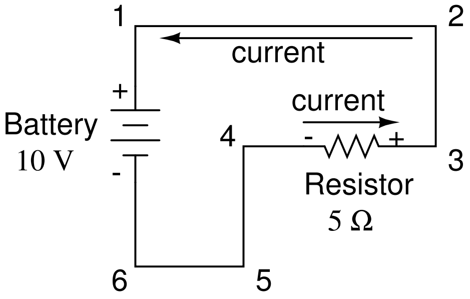
If those points are *supposed* to be electrically common but they register otherwise, then I know that there is an "open failure" between those points.

One final note: for most practical purposes, wire conductors can be assumed to possess zero resistance from end to end. In reality, however, there will always be some small amount of resistance encountered along the length of a wire, unless it's a superconducting wire. Knowing this, we need to bear in mind that the principles learned here about electrically common points are all valid to a large degree, but not to an *absolute* degree. That is, the rule that electrically common points are guaranteed to have zero voltage between them is more accurately stated as such: electrically common points will have *very little* voltage dropped between them. That small, virtually unavoidable trace of resistance found in any piece of connecting wire is bound to create a small voltage across the length of it as current is conducted through. So long as you understand that these rules are based upon *ideal* conditions, you won't be perplexed when you come across some condition appearing to be an exception to the rule.

14.15 Polarity of voltage drops

We can trace the direction that electrons will flow in the same circuit by starting at the negative (-) terminal and following through to the positive (+) terminal of the battery, the only source of voltage in the circuit. From this we can see that the electrons are moving counter-clockwise, from point 6 to 5 to 4 to 3 to 2 to 1 and back to 6 again.

As the current encounters the 5 Ω resistance, voltage is dropped across the resistor's ends. The signs of this voltage drop is negative (-) at point 4 with respect to positive (+) at point 3. We can mark the polarity of the resistor's voltage drop with these negative and positive symbols, in accordance with the direction of current (whichever end of the resistor the current is *entering* is negative with respect to the end of the resistor it is *exiting*:



We could make our table of voltages a little more complete by marking the polarity of the voltage for each pair of points in this circuit:

```
Between points 1 (+) and 4 (-) = 10 volts
Between points 2 (+) and 4 (-) = 10 volts
Between points 3 (+) and 4 (-) = 10 volts
Between points 1 (+) and 5 (-) = 10 volts
Between points 2 (+) and 5 (-) = 10 volts
Between points 3 (+) and 5 (-) = 10 volts
Between points 1 (+) and 6 (-) = 10 volts
Between points 2 (+) and 6 (-) = 10 volts
Between points 3 (+) and 6 (-) = 10 volts
```

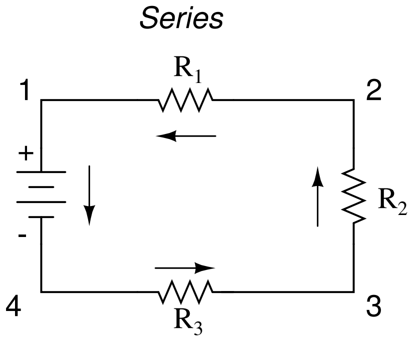
While it might seem a little silly to document polarity of voltage drop in this circuit, it is an important concept to master. It will be critically important in the analysis of more complex circuits involving multiple resistors and/or batteries.

It should be understood that polarity has nothing to do with Ohm's Law: there will never be negative voltages, currents, or resistance entered into any Ohm's Law equations! There are other mathematical principles of electricity that do take polarity into account through the use of signs (+ or -), but not Ohm's Law.

14.16 What are "series" and "parallel" circuits?

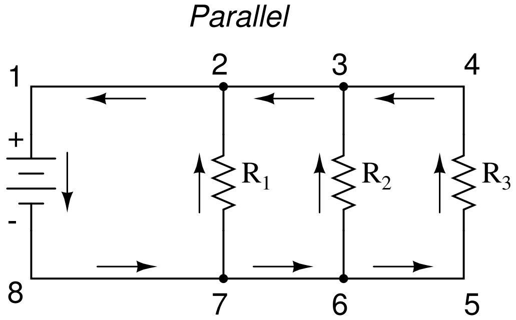
Circuits consisting of just one battery and one load resistance are very simple to analyze, but they are not often found in practical applications. Usually, we find circuits where more than two components are connected together.

There are two basic ways in which to connect more than two circuit components: *series* and *parallel*. First, an example of a series circuit:



Here, we have three resistors (labeled R_1, R_2, and R_3), connected in a long chain from one terminal of the battery to the other. (It should be noted that the subscript labeling – those little numbers to the lower-right of the letter "R" – are unrelated to the resistor values in ohms. They serve only to identify one resistor from another.) The defining characteristic of a series circuit is that there is only one path for electrons to flow. In this circuit the electrons flow in a counter-clockwise direction, from point 4 to point 3 to point 2 to point 1 and back around to 4.

Now, let's look at the other type of circuit, a parallel configuration:

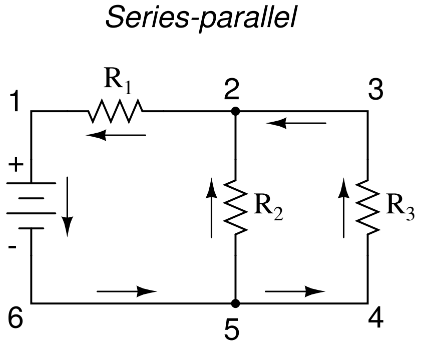


Again, we have three resistors, but this time they form more than one continuous path for electrons to flow. There's one path from 8 to 7 to 2 to 1 and back to 8 again. There's another from 8 to 7 to 6 to 3 to 2 to 1 and back to 8 again. And then there's a third path from 8 to 7 to

6 to 5 to 4 to 3 to 2 to 1 and back to 8 again. Each individual path (through R_1, R_2, and R_3) is called a *branch.*

The defining characteristic of a parallel circuit is that all components are connected between the same set of electrically common points. Looking at the schematic diagram, we see that points 1, 2, 3, and 4 are all electrically common. So are points 8, 7, 6, and 5. Note that all resistors as well as the battery are connected between these two sets of points.

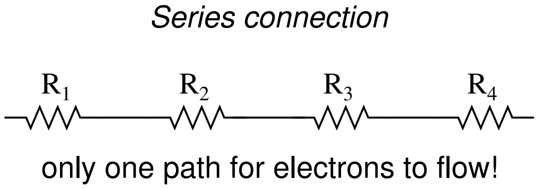
And, of course, the complexity doesn't stop at simple series and parallel either! We can have circuits that are a combination of series and parallel, too:



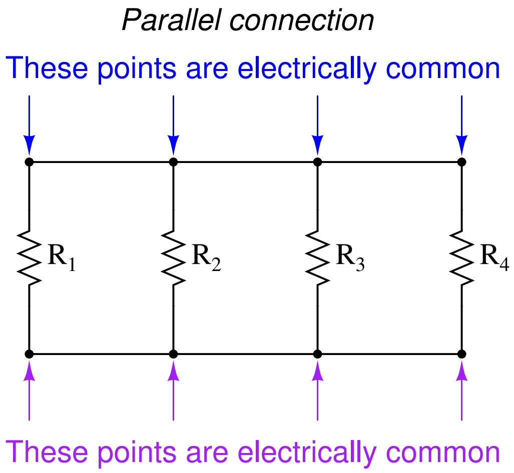
In this circuit, we have two loops for electrons to flow through: one from 6 to 5 to 2 to 1 and back to 6 again, and another from 6 to 5 to 4 to 3 to 2 to 1 and back to 6 again. Notice how both current paths go through R_1 (from point 2 to point 1). In this configuration, we'd say that R_2 and R_3 are in parallel with each other, while R_1 is in series with the parallel combination of R_2 and R_3.

This is just a preview of things to come. Don't worry! We'll explore all these circuit configurations in detail, one at a time!

The basic idea of a "series" connection is that components are connected end-to-end in a line to form a single path for electrons to flow:



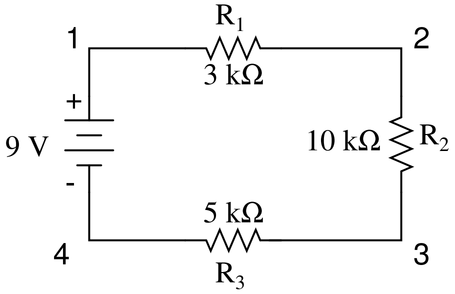
The basic idea of a "parallel" connection, on the other hand, is that all components are connected across each other's leads. In a purely parallel circuit, there are never more than two sets of electrically common points, no matter how many components are connected. There are many paths for electrons to flow, but only one voltage across all components:


Parallel connection
These points are electrically common
R1
R2
R3
R4
These points are electrically common

Series and parallel resistor configurations have very different electrical properties. We'll explore the properties of each configuration in the sections to come.

14.17 Simple series circuits

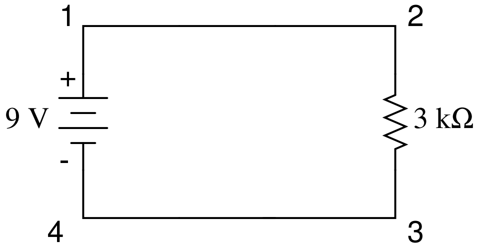
Let's start with a series circuit consisting of three resistors and a single battery:



The first principle to understand about series circuits is that the amount of current is the same through any component in the circuit. This is because there is only one path for electrons to flow in a series circuit, and because free electrons flow through conductors like marbles in a tube, the rate of flow (marble speed) at any point in the circuit (tube) at any specific point in time must be equal.

From the way that the 9 volt battery is arranged, we can tell that the electrons in this circuit will flow in a counter-clockwise direction, from point 4 to 3 to 2 to 1 and back to 4. However, we have one source of voltage and three resistances. How do we use Ohm's Law here?

An important caveat to Ohm's Law is that all quantities (voltage, current, resistance, and power) must relate to each other in terms of the same two points in a circuit. For instance, with a single-battery, single-resistor circuit, we could easily calculate any quantity because they all applied to the same two points in the circuit:

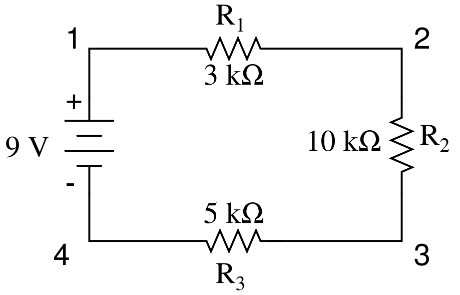


$$I = \frac{E}{R}$$

$$I = \frac{9 \text{ volts}}{3 \text{ k}\Omega} = 3 \text{ mA}$$

Since points 1 and 2 are connected together with wire of negligible resistance, as are points 3 and 4, we can say that point 1 is electrically common to point 2, and that point 3 is electrically common to point 4. Since we know we have 9 volts of electromotive force between points 1 and 4 (directly across the battery), and since point 2 is common to point 1 and point 3 common to point 4, we must also have 9 volts between points 2 and 3 (directly across the resistor). Therefore, we can apply Ohm's Law (I = E/R) to the current through the resistor, because we know the voltage (E) across the resistor and the resistance (R) of that resistor. All terms (E, I, R) apply to the same two points in the circuit, to that same resistor, so we can use the Ohm's Law formula with no reservation.

However, in circuits containing more than one resistor, we must be careful in how we apply Ohm's Law. In the three-resistor example circuit below, we know that we have 9 volts between points 1 and 4, which is the amount of electromotive force trying to push electrons through the series combination of R_1, R_2, and R_3. However, we cannot take the value of 9 volts and divide it by 3k, 10k or 5k Ω to try to find a current value, because we don't know how much voltage is across any one of those resistors, individually.



The figure of 9 volts is a *total* quantity for the whole circuit, whereas the figures of 3k, 10k, and 5k Ω are *individual* quantities for individual resistors. If we were to plug a figure for total voltage into an Ohm's Law equation with a figure for individual resistance, the result would not relate accurately to any quantity in the real circuit.

For R_1, Ohm's Law will relate the amount of voltage across R_1 with the current through R_1, given R_1's resistance, 3kΩ:

$$I_{R1} = \frac{E_{R1}}{3\ \text{k}\Omega} \qquad E_{R1} = I_{R1}(3\ \text{k}\Omega)$$

But, since we don't know the voltage across R_1 (only the total voltage supplied by the battery across the three-resistor series combination) and we don't know the current through R_1, we can't do any calculations with either formula. The same goes for R_2 and R_3: we can apply the Ohm's Law equations if and only if all terms are representative of their respective quantities between the same two points in the circuit.

So what can we do? We know the voltage of the source (9 volts) applied across the series combination of R_1, R_2, and R_3, and we know the resistances of each resistor, but since those quantities aren't in the same context, we can't use Ohm's Law to determine the circuit current. If only we knew what the *total* resistance was for the circuit: then we could calculate *total* current with our figure for *total* voltage (I=E/R).

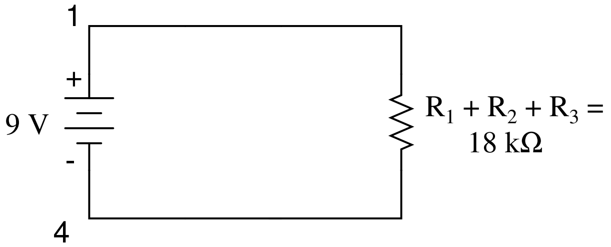
This brings us to the second principle of series circuits: the total resistance of any series circuit is equal to the sum of the individual resistances. This should make intuitive sense: the more resistors in series that the electrons must flow through, the more difficult it will be for those electrons to flow. In the example problem, we had a 3 kΩ, 10 kΩ, and 5 kΩ resistor in series, giving us a total resistance of 18 kΩ:

$$R_{total} = R_1 + R_2 + R_3$$

$$R_{total} = 3\ \text{k}\Omega + 10\ \text{k}\Omega + 5\ \text{k}\Omega$$

$$R_{total} = 18\ \text{k}\Omega$$

In essence, we've calculated the equivalent resistance of R_1, R_2, and R_3 combined. Knowing this, we could re-draw the circuit with a single equivalent resistor representing the series combination of R_1, R_2, and R_3:

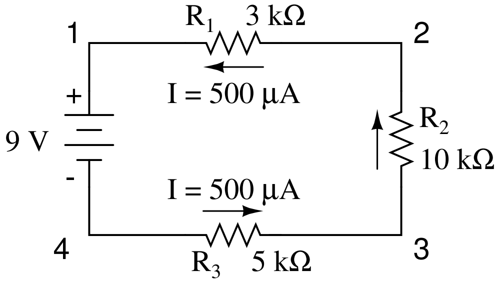


Now we have all the necessary information to calculate circuit current, because we have the voltage between points 1 and 4 (9 volts) and the resistance between points 1 and 4 (18 kΩ):

$$I_{total} = \frac{E_{total}}{R_{total}}$$

$$I_{total} = \frac{9\ \text{volts}}{18\ \text{k}\Omega} = 500\ \mu\text{A}$$

Knowing that current is equal through all components of a series circuit (and we just determined the current through the battery), we can go back to our original circuit schematic and note the current through each component:



Now that we know the amount of current through each resistor, we can use Ohm's Law to determine the voltage drop across each one (applying Ohm's Law in its proper context):

$$E_{R1} = I_{R1}\, R_1 \qquad E_{R2} = I_{R2}\, R_2 \qquad E_{R3} = I_{R3}\, R_3$$

$$E_{R1} = (500\ \mu\text{A})(3\ \text{k}\Omega) = 1.5\ \text{V}$$

$$E_{R2} = (500\ \mu\text{A})(10\ \text{k}\Omega) = 5\ \text{V}$$

$$E_{R3} = (500\ \mu\text{A})(5\ \text{k}\Omega) = 2.5\ \text{V}$$

Notice the voltage drops across each resistor, and how the sum of the voltage drops (1.5 + 5 + 2.5) is equal to the battery (supply) voltage: 9 volts. This is the third principle of series circuits: that the supply voltage is equal to the sum of the individual voltage drops.

However, the method we just used to analyze this simple series circuit can be streamlined for better understanding. By using a table to list all voltages, currents, and resistances in the circuit, it becomes very easy to see which of those quantities can be properly related in any Ohm's Law equation:

	R_1	R_2	R_3	Total	
E					Volts
I					Amps
R					Ohms
	↑ *Ohm's Law*	↑ *Ohm's Law*	↑ *Ohm's Law*	↑ *Ohm's Law*	

The rule with such a table is to apply Ohm's Law only to the values within each vertical column. For instance, E_{R1} only with I_{R1} and R_1; E_{R2} only with I_{R2} and R_2; etc. You begin your analysis by filling in those elements of the table that are given to you from the beginning:

	R_1	R_2	R_3	Total	
E				**9**	Volts
I					Amps
R	**3k**	**10k**	**5k**		Ohms

As you can see from the arrangement of the data, we can't apply the 9 volts of E_T (total voltage) to any of the resistances (R_1, R_2, or R_3) in any Ohm's Law formula because they're in different columns. The 9 volts of battery voltage is *not* applied directly across R_1, R_2, or R_3. However, we can use our "rules" of series circuits to fill in blank spots on a horizontal row. In this case, we can use the series rule of resistances to determine a total resistance from the *sum* of individual resistances:

	R_1	R_2	R_3	Total	
E				9	Volts
I					Amps
R	3k	10k	5k	**18k**	Ohms

Rule of series circuits

$R_T = R_1 + R_2 + R_3$

Now, with a value for total resistance inserted into the rightmost ("Total") column, we can apply Ohm's Law of I=E/R to total voltage and total resistance to arrive at a total current of 500 μA:

	R_1	R_2	R_3	Total	
E				9	Volts
I				**500µ**	Amps
R	3k	10k	5k	18k	Ohms

Ohm's Law

Then, knowing that the current is shared equally by all components of a series circuit (another "rule" of series circuits), we can fill in the currents for each resistor from the current figure just calculated:

	R_1	R_2	R_3	Total	
E				9	Volts
I	**500µ**	**500µ**	**500µ**	**500µ**	Amps
R	3k	10k	5k	18k	Ohms

Rule of series circuits

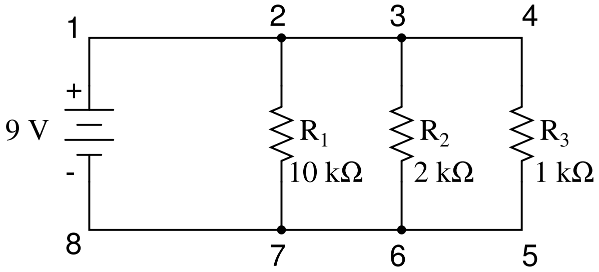
$I_T = I_1 = I_2 = I_3$

Finally, we can use Ohm's Law to determine the voltage drop across each resistor, one column at a time:

	R_1	R_2	R_3	Total	
E	**1.5**	**5**	**2.5**	9	Volts
I	500µ	500µ	500µ	500µ	Amps
R	3k	10k	5k	18k	Ohms
	↑ *Ohm's Law*	↑ *Ohm's Law*	↑ *Ohm's Law*		

14.18 Simple parallel circuits

Let's start with a parallel circuit consisting of three resistors and a single battery:



The first principle to understand about parallel circuits is that the voltage is equal across all components in the circuit. This is because there are only two sets of electrically common points in a parallel circuit, and voltage measured between sets of common points must always be the same at any given time. Therefore, in the above circuit, the voltage across R_1 is equal to the voltage across R_2 which is equal to the voltage across R_3 which is equal to the voltage across the battery. This equality of voltages can be represented in another table for our starting values:

	R_1	R_2	R_3	Total	
E	9	9	9	9	Volts
I					Amps
R	10k	2k	1k		Ohms

Just as in the case of series circuits, the same caveat for Ohm's Law applies: values for voltage, current, and resistance must be in the same context in order for the calculations to work correctly. However, in the above example circuit, we can immediately apply Ohm's Law to each resistor to find its current because we know the voltage across each resistor (9 volts) and the resistance of each resistor:

$$I_{R1} = \frac{E_{R1}}{R_1} \qquad I_{R2} = \frac{E_{R2}}{R_2} \qquad I_{R3} = \frac{E_{R3}}{R_3}$$

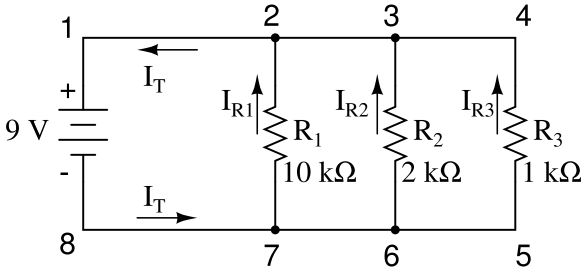
$$I_{R1} = \frac{9\ \text{V}}{10\ \text{k}\Omega} = 0.9\ \text{mA}$$

$$I_{R2} = \frac{9\ \text{V}}{2\ \text{k}\Omega} = 4.5\ \text{mA}$$

$$I_{R3} = \frac{9\ \text{V}}{1\ \text{k}\Omega} = 9\ \text{mA}$$

	R_1	R_2	R_3	Total	
E	9	9	9	9	Volts
I	**0.9m**	**4.5m**	**9m**		Amps
R	10k	2k	1k		Ohms
	↑ *Ohm's Law*	↑ *Ohm's Law*	↑ *Ohm's Law*		

At this point we still don't know what the total current or total resistance for this parallel circuit is, so we can't apply Ohm's Law to the rightmost (”Total”) column. However, if we think carefully about what is happening it should become apparent that the total current must equal the sum of all individual resistor (”branch”) currents:



As the total current exits the negative (-) battery terminal at point 8 and travels through the circuit, some of the flow splits off at point 7 to go up through R_1, some more splits off at point 6 to go up through R_2, and the remainder goes up through R_3. Like a river branching into several smaller streams, the combined flow rates of all streams must equal the flow rate of the whole river. The same thing is encountered where the currents through R_1, R_2, and R_3 join to flow back to the positive terminal of the battery (+) toward point 1: the flow of electrons from point 2 to point 1 must equal the sum of the (branch) currents through R_1, R_2, and R_3.

This is the second principle of parallel circuits: the total circuit current is equal to the sum of the individual branch currents. Using this principle, we can fill in the I_T spot on our table with the sum of I_{R1}, I_{R2}, and I_{R3}:

	R_1	R_2	R_3	Total	
E	9	9	9	9	Volts
I	0.9m	4.5m	9m	**14.4m**	Amps
R	10k	2k	1k		Ohms

Rule of parallel circuits

$$I_{total} = I_1 + I_2 + I_3$$

Finally, applying Ohm's Law to the rightmost (”Total”) column, we can calculate the total circuit resistance:

	R_1	R_2	R_3	Total	
E	9	9	9	9	Volts
I	0.9m	4.5m	9m	14.4m	Amps
R	10k	2k	1k	**625**	Ohms

$$R_{total} = \frac{E_{total}}{I_{total}} = \frac{9 \text{ V}}{14.4 \text{ mA}} = 625 \ \Omega$$

Ohm's Law

Please note something very important here. The total circuit resistance is only 625 Ω: *less* than any one of the individual resistors. In the series circuit, where the total resistance was the sum of the individual resistances, the total was bound to be *greater* than any one of the resistors individually. Here in the parallel circuit, however, the opposite is true: we say that the individual resistances *diminish* rather than *add* to make the total. This principle completes our triad of ”rules” for parallel circuits, just as series circuits were found to have three rules for voltage, current, and resistance. Mathematically, the relationship between total resistance and individual resistances in a parallel circuit looks like this:

$$R_{total} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}}$$

The same basic form of equation works for *any* number of resistors connected together in parallel, just add as many 1/R terms on the denominator of the fraction as needed to accommodate all parallel resistors in the circuit.

14.19 Power calculations

When calculating the power dissipation of resistive components, use any one of the three power equations to derive and answer from values of voltage, current, and/or resistance pertaining to each component:

Power equations

$$P = IE \qquad P = \frac{E^2}{R} \qquad P = I^2R$$

This is easily managed by adding another row to our familiar table of voltages, currents, and resistances:

	R_1	R_2	R_3	Total	
E					Volts
I					Amps
R					Ohms
P					Watts

Power for any particular table column can be found by the appropriate Ohm's Law equation (*appropriate* based on what figures are present for E, I, and R in that column).

An interesting rule for total power versus individual power is that it is additive for *any* configuration of circuit: series, parallel, series/parallel, or otherwise. Power is a measure of rate of work, and since power dissipated *must* equal the total power applied by the source(s) (as per the Law of Conservation of Energy in physics), circuit configuration has no effect on the mathematics.

14.20 Correct use of Ohm's Law

When working through worked examples it is important to try to figure out what you are doing correctly as well as what you are doing wrong. Make sure you don't stop doing the good things and try to correct the mistakes.

Circuit questions form a large part of high school and early university courses and it is important to understand the concepts properly. One common mistake which students make we'll discuss here so you know to look out for it when you are working through examples and studying.

When applying Ohm's Laws students often mix up the contexts of voltage, current, and resistance. This means a student might mistakenly use a value for I through one resistor and the value for V across another resistor or a set of connected resistors.

Remember this important rule: The variables used in Ohm's Law equations must be *common* to the same two points in the circuit under consideration. This is especially important in series-parallel combination circuits where nearby components may have different values for both voltage drop *and* current.

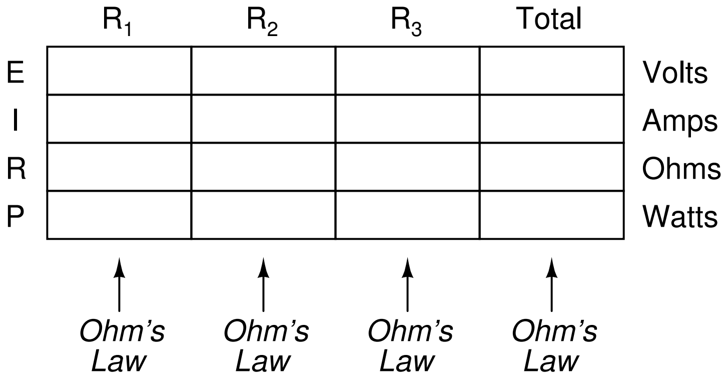
When using Ohm's Law to calculate a variable for a single component:

- be sure the voltage you're using is solely across that single component **and**
- the current you're referencing is solely through that single component **and**
- the resistance you're referencing is solely for that single component.

When calculating a variable for a set of components in a circuit, be sure that the voltage, current, and resistance values are specific to that complete set of components only!

A good way to remember this is to pay close attention to the *two points* on either side of the component or set of components. Making sure that the voltage in question is across those two points, that the current in question is the electron flow from one of those points all the way to the other point, that the resistance in question is the equivalent of a single resistor between those two points, and that the power in question is the total power dissipated by all components between those two points.

The "table" method presented for both series and parallel circuits in this chapter is a way to keep the components correct when using Ohm's Law. In a table like the one shown below, you are only allowed to apply an Ohm's Law equation for the values of a single *vertical* column at a time:

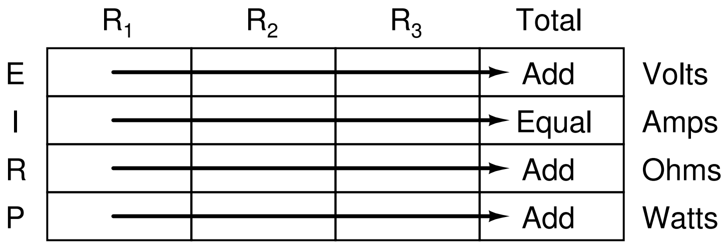


	R_1	R_2	R_3	Total	
E					Volts
I					Amps
R					Ohms
P					Watts

Ohm's Law
Ohm's Law
Ohm's Law
Ohm's Law

Deriving values *horizontally* across columns is allowable as per the principles of series and parallel circuits:

For series circuits:



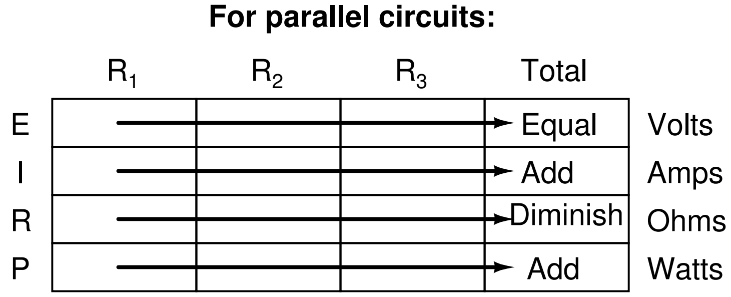
	R_1	R_2	R_3	Total	
E				Add	Volts
I				Equal	Amps
R				Add	Ohms
P				Add	Watts

$$E_{total} = E_1 + E_2 + E_3$$

$$I_{total} = I_1 = I_2 = I_3$$

$$R_{total} = R_1 + R_2 + R_3$$

$$P_{total} = P_1 + P_2 + P_3$$



For parallel circuits:

	R_1	R_2	R_3	Total	
E				→ Equal	Volts
I				→ Add	Amps
R				→ Diminish	Ohms
P				→ Add	Watts

$$E_{total} = E_1 = E_2 = E_3$$

$$I_{total} = I_1 + I_2 + I_3$$

$$R_{total} = \frac{1}{\frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3}}$$

$$P_{total} = P_1 + P_2 + P_3$$

The "table" method helps to keep track of all relevant quantities. It also facilitates cross-checking of answers by making it easy to solve for the original unknown variables through other methods, or by working backwards to solve for the initially given values from your solutions.

For example, if you have just solved for all unknown voltages, currents, and resistances in a circuit, you can check your work by adding a row at the bottom for power calculations on each resistor, seeing whether or not all the individual power values add up to the total power. If not, then you must have made a mistake somewhere! While this technique of "cross-checking" your work is nothing new, using the table to arrange all the data for the cross-check(s) results in a minimum of confusion.

Aside: Although checking your work might not be fun when you have just worked hard on the problem the benefits are great. Coming back to a problem after a small break (trying another problem) often helps to find simple mistakes. If you have done all the work then finding a simple mistake will be quick to fix because you know exactly what you need to do. Also if you start finding mistakes while checking you'll build a mental list and find that you'll stop making them after a while.
Do it and you'll find it will pay off!

14.21 Conductor size

The width of a conductor affects the flow of electrons through it. The broader the cross-sectional area (thickness or area of a sl) of the conductor, the more room for electrons to flow, and consequently, the easier it is for flow to occur (less resistance).

Electrical wire is usually round in cross-section (although there are some unique exceptions to this rule), and comes in two basic varieties: solid and stranded. Solid copper wire is just as it sounds: a single, solid strand of copper the whole length of the wire. Stranded wire is composed of smaller strands of solid copper wire twisted together to form a single, larger conductor. The greatest benefit of stranded wire is its mechanical flexibility, being able to withstand repeated bending and twisting much better than solid copper (which tends to fatigue and break after time).

14.22 Fuses

Normally, the ampacity rating of a conductor is a circuit design limit never to be intentionally exceeded, but there is an application where ampacity exceedence is expected: in the case of *fuses.*

A fuse is nothing more than a short length of wire designed to melt and separate in the event of excessive current. Fuses are always connected in series with the component(s) to be protected from overcurrent, so that when the fuse *blows* (opens) it will open the entire circuit and stop current through the component(s). A fuse connected in one branch of a parallel circuit, of course, would not affect current through any of the other branches.

Normally, the thin piece of fuse wire is contained within a safety sheath to minimize hazards of arc blast if the wire burns open with violent force, as can happen in the case of severe overcurrents. In the case of small automotive fuses, the sheath is transparent so that the fusible element can be visually inspected. Residential wiring used to commonly employ screw-in fuses with glass bodies and a thin, narrow metal foil strip in the middle.

14.23 Important Equations and Quantities

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
			or	

Table 14.1: Units used in **Electricity and Magnetism**

- **REVIEW:**
- A *circuit* is an unbroken loop of conductive material that allows electrons to flow through continuously without beginning or end.
- If a circuit is "broken," that means it's conductive elements no longer form a complete path, and continuous electron flow cannot occur in it.
- The location of a break in a circuit is irrelevant to its inability to sustain continuous electron flow. *Any* break, *anywhere* in a circuit prevents electron flow throughout the circuit.
- **REVIEW:**
- Electrons can be motivated to flow through a conductor by a the same force manifested in static electricity.
- *Voltage* is the measure of specific potential energy (potential energy per unit charge) between two locations. In layman's terms, it is the measure of "push" available to motivate electrons.
- Voltage, as an expression of potential energy, is always relative between two locations, or points. Sometimes it is called a voltage "drop."

- When a voltage source is connected to a circuit, the voltage will cause a uniform flow of electrons through that circuit called a *current.*
- In a single (one loop) circuit, the amount current of current at any point is the same as the amount of current at any other point.
- If a circuit containing a voltage source is broken, the full voltage of that source will appear across the points of the break.
- The +/- orientation a voltage drop is called the *polarity.* It is also relative between two points.
- **REVIEW:**
- *Resistance* is the measure of opposition to electric current.
- A *short circuit* is an electric circuit offering little or no resistance to the flow of electrons. Short circuits are dangerous with high voltage power sources because the high currents encountered can cause large amounts of heat energy to be released.
- An *open circuit* is one where the continuity has been broken by an interruption in the path for electrons to flow.
- A *closed circuit* is one that is complete, with good continuity throughout.
- A device designed to open or close a circuit under controlled conditions is called a *switch.*
- The terms *"open"* and *"closed"* refer to switches as well as entire circuits. An open switch is one without continuity: electrons cannot flow through it. A closed switch is one that provides a direct (low resistance) path for electrons to flow through.
- **REVIEW:**
- Connecting wires in a circuit are assumed to have zero resistance unless otherwise stated.
- Wires in a circuit can be shortened or lengthened without impacting the circuit's function – all that matters is that the components are attached to one another in the same sequence.
- Points directly connected together in a circuit by zero resistance (wire) are considered to be *electrically common.*
- Electrically common points, with zero resistance between them, will have zero voltage dropped between them, regardless of the magnitude of current (ideally).
- The voltage or resistance readings referenced between sets of electrically common points will be the same.
- These rules apply to *ideal* conditions, where connecting wires are assumed to possess absolutely zero resistance. In real life this will probably not be the case, but wire resistances should be low enough so that the general principles stated here still hold.
- **REVIEW:**
- Power is additive in *any* configuration of resistive circuit: $\mathrm{P}_{Total} = \mathrm{P}_1 + \mathrm{P}_2 + \ldots \mathrm{P}_n$
- **REVIEW:**

- When electrons flow through a conductor, a magnetic field will be produced around that conductor.
- The left-hand rule states that the magnetic flux lines produced by a current-carrying wire will be oriented the same direction as the curled fingers of a person's left hand (in the "hitchhiking" position), with the thumb pointing in the direction of electron flow.
- The magnetic field force produced by a current-carrying wire can be greatly increased by shaping the wire into a coil instead of a straight line. If wound in a coil shape, the magnetic field will be oriented along the axis of the coil's length.
- The magnetic field force produced by an electromagnet (called the *magnetomotive force*, or mmf), is proportional to the product (multiplication) of the current through the electromagnet and the number of complete coil "turns" formed by the wire.

Chapter 15

Magnets and Electromagnetism

sectionPermanent magnets

Magnetism has been known to mankind for many thousands of years. *Lodestone*, a magnetized form of the iron oxide mineral *magnetite* which has the property of attracting iron objects, is referred to in old European and Asian historical records, around 800 BC in Europe and earlier in the East, around 2600 BC. The root of the English word magnet is the Greek word magnes, thought to be derived from Magnesia in Asia Minor, once an important source of lodestone.

Lodestone was used as a navigational compass as it was found to orient itself in a north-south direction if left free to rotate by suspension on a string or on a float in water.

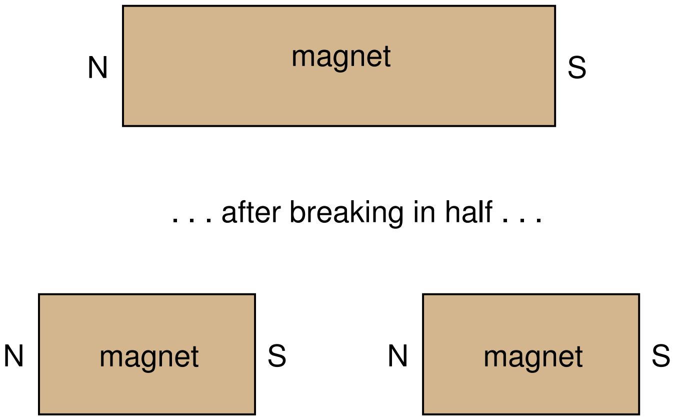
Interesting Fact:

A compass is a navigational instrument for finding directions. It consists of a magnetised pointer free to align itself accurately with Earth's magnetic field. A compass provides a known reference direction which is of great assistance in navigation. The cardinal points are north, south, east and west. A compass can be used in conjunction with a clock and a sextant to provide a very accurate navigation capability. This device greatly improved maritime trade by making travel safer and more efficient.

A compass can be any magnetic device using a needle to indicate the direction of the magnetic north of a planet's magnetosphere. Any instrument with a magnetized bar or needle turning freely upon a pivot and pointing in a northerly and southerly direction can be considered a compass.

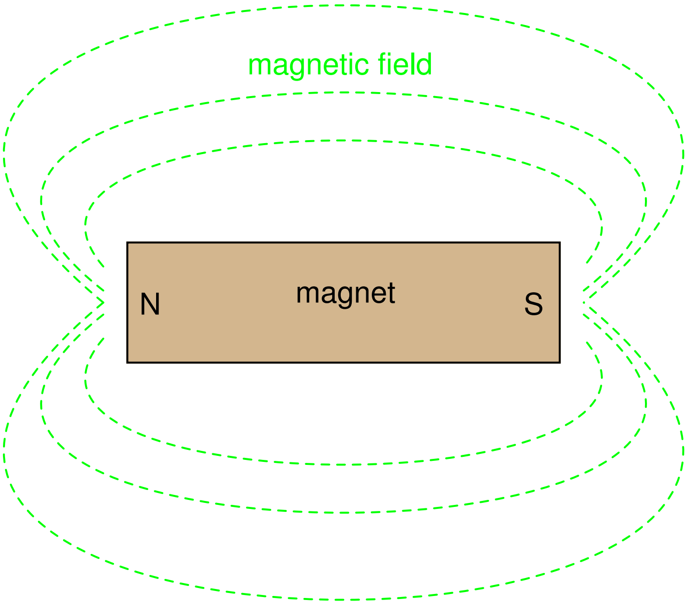
Aside: In 1269, Frenchmen Peter Peregrinus and Pierre de Maricourt, using a compass and a lodestone, found that the magnetic force of the lodestone was different at the opposite ends, which they defined to be the poles of the magnet.

Like poles of magnets repel one another whilst unlike poles attract. These poles always occur in pairs. It is impossible to isolate a single pole. Breaking a piece of magnet in half results in two pieces, each with it's own pair of poles.



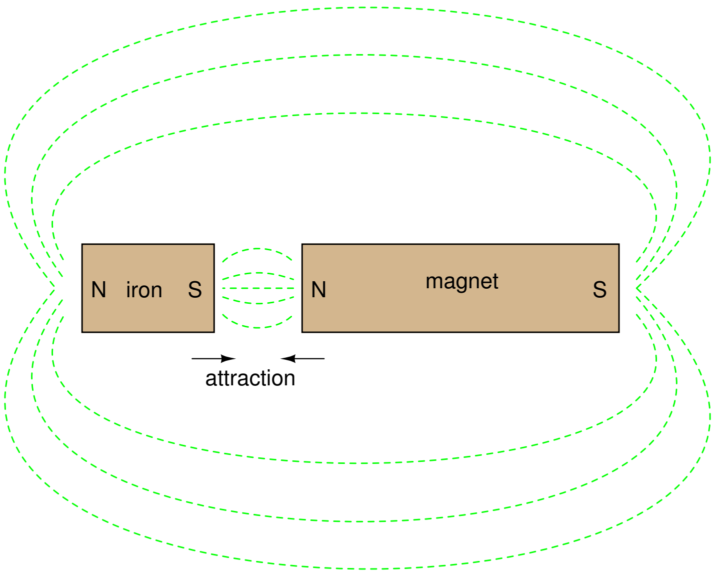
The Earth itself is a magnet. Its magnetic poles are approximately aligned along the Earth's axis of rotation. The magnitude of forces between the poles of magnets follows an inverse square law; *i. e.* it varies inversely as the square of the distance of separation.

Magnetic forces are a result of magnetic fields. By placing a magnet underneath a piece of paper and sprinkling iron filings on top one can map the magnetic field. The filings align themselves parallel to the field. Magnetic fields can be represented by magnetic field lines which are parallel to the magnetic field and whose spacing represents the relative strength of the magnetic field. The strength of the magnetic field is referred to as the magnetic flux. Magnetic field lines form closed loops. In a bar magnet magnetic field lines emerge at one pole and then curve around to the other pole with the rest of the loop being inside the magnet.



As already said, opposite poles of a magnet attract each other and bringing them together results in their magnetic field lines converging. Like poles of a magnet repel each other and bringing them together results in their magnetic field lines diverging.

Ferromagnetism is a phenomenon exhibited by materials like iron, nickel or cobalt. These materials are known as permanent magnets. They always magnetize so as to be attracted to a magnet, regardless of which magnetic pole is brought toward the unmagnetized iron:



Interesting Fact:

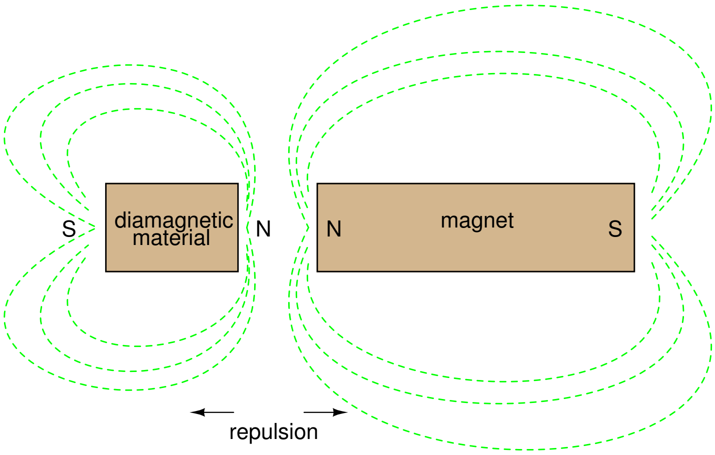
The cause of Earth's magnetic field is not known for certain, but is possibly explained by the dynamo theory. The magnetic field extends several tens of thousands of kilometers into space. The field is approximately a magnetic dipole, with one pole near the geographic north pole and the other near the geographic south pole. An imaginary line joining the magnetic poles would be inclined by approximately 11.3 from the planet's axis of rotation. The location of the magnetic poles is not static but wanders as much as several kilometers a year. The two poles wander independently of each other and are not at exact opposite positions on the globe.

The Earth's magnetic field reverses at intervals, ranging from tens of thousands to many millions of years, with an average interval of 250,000 years. It is believed that this last occurred some 780,000 years ago. The mechanism responsible for geomagnetic reversals is not well understood. When the North reappears in the opposite direction, we would interpret this as a reversal, whereas turning off and returning in the same direction is called a geomagnetic excursion. At present, the overall geomagnetic field is becoming weaker at a rate which would, if it continues, cause the field to disappear, albeit temporarily, by about around 3000-4000 AD. The deterioration began roughly 150 years ago and has accelerated in the past several years. So far the strength of the earth's field has decreased by 10 to 15 percent.

The ability of a ferromagnetic material tends to retain its magnetization after an external field is removed is called it's *retentivity*.

Paramagnetic materials are materials like aluminum or platinum which become magnetized in an external magnetic field in a similar way to ferromagnetic materials but lose their magnetism when the external magnetic field is removed.

Diamagnetism is exhibited by materials like copper or bismuth which become magnetized in a magnetic field with a polarity opposite to the external magnetic field. Unlike iron, they are slightly repelled by a magnet



The cause of Earth's magnetic field is not known for certain, but is possibly explained by the dynamo theory. The magnetic field extends several tens of thousands of kilometers into space. The field is approximately a magnetic dipole, with one pole near the geographic north pole and the other near the geographic south pole. An imaginary line joining the magnetic poles would be inclined by approximately 11.3 from the planet's axis of rotation. The location of the magnetic poles is not static but wanders as much as several kilometers a year. The two poles wander independently of each other and are not at exact opposite positions on the globe. Currently the south magnetic pole is further from the geographic south pole than than north magnetic pole is from the north geographic pole.

The strength of the field at the Earth's surface at this time ranges from less than 30 microtesla (0.3 gauss) in an area including most of South America and South Africa to over 60 microtesla (0.6 gauss) around the magnetic poles in northern Canada and south of Australia, and in part of Siberia. The field is similar to that of a bar magnet, but this similarity is superficial. The magnetic field of a bar magnet, or any other type of permanent magnet, is created by the coordinated motions of electrons (negatively charged particles) within iron atoms. The Earth's core, however, is hotter than 1043 K, the temperature at which the orientations of electron orbits within iron become randomized. Therefore the Earth's magnetic field is not caused by magnetised iron deposits, but mostly by electric currents (known as telluric currents). Another feature that distinguishes the Earth magnetically from a bar magnet is its magnetosphere.

A magnetosphere is the region around an astronomical object, in which phenomena are dominated by its magnetic field. Earth is surrounded by a magnetosphere, as are the magnetized planets Jupiter, Saturn, Uranus and Neptune. Mercury is magnetized, but too weakly to trap plasma. Mars has patchy surface magnetization.

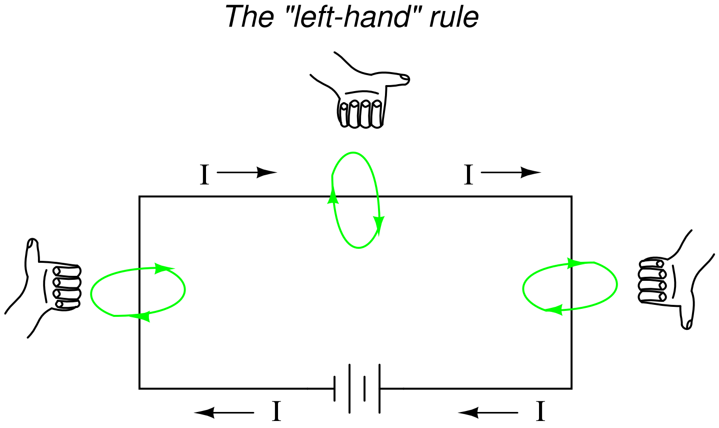
The distant field of Earth is greatly modified by the solar wind, a hot outflow from the sun, consisting of solar ions (mainly hydrogen) moving at about 400 km/s . Earth's magnetic field forms an obstacle to the solar wind.

The Earth's magnetic field reverses at intervals, ranging from tens of thousands to many millions of years, with an average interval of 250,000 years. It is believed that this last occurred some 780,000 years ago. The mechanism responsible for geomagnetic reversals is not well understood. When the North reappears in the opposite direction, we would interpret this as a reversal, whereas turning off and returning in the same direction is called a geomagnetic excursion. At present, the overall geomagnetic field is becoming weaker at a rate which would, if it continues, cause the field to disappear, albeit temporarily, by about around 3000-4000 AD. The deterioration began roughly 150 years ago and has accelerated in the past several years. So far the strength of the earth's field has decreased by 10 to 15 percent.

15.1 Electromagnetism

> The discovery of the relationship between magnetism and electricity was, like so many other scientific discoveries, stumbled upon almost by accident. The Danish physicist Hans Christian Oersted was lecturing one day in 1820 on the *possibility* of electricity and magnetism being related to one another, and in the process demonstrated it conclusively by experiment in front of his whole class! By passing an electric current through a metal wire suspended above a magnetic compass, Oersted was able to produce a definite motion of the compass needle in response to the current. What began as conjecture at the start of the class session was confirmed as fact at the end. Needless to say, Oersted had to revise his lecture notes for future classes! His serendipitous discovery paved the way for a whole new branch of science: electromagnetics.

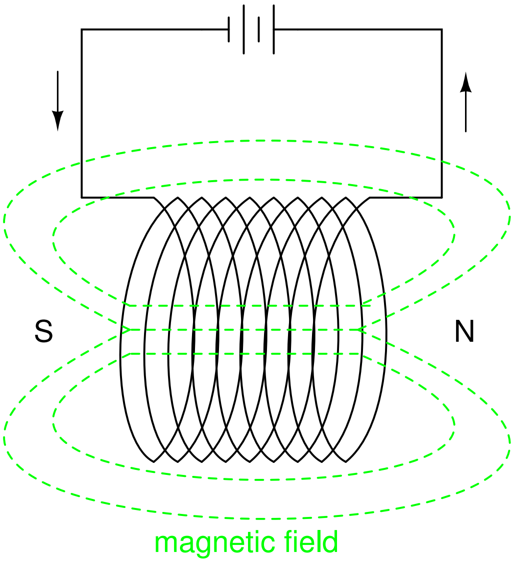
Detailed experiments showed that the magnetic field produced by an electric current is always oriented perpendicular to the direction of flow. A simple method of showing this relationship is called the *left-hand rule*. Simply stated, the left-hand rule says that the magnetic flux lines produced by a current-carrying wire will be oriented the same direction as the curled fingers of a person's left hand (in the "hitchhiking" position), with the thumb pointing in the direction of electron flow:



The magnetic field encircles this straight piece of current-carrying wire, the magnetic flux lines having no definite "north" or "south' poles.

(NOTE TO SELF: Need to add wires attracting or wires repelling)

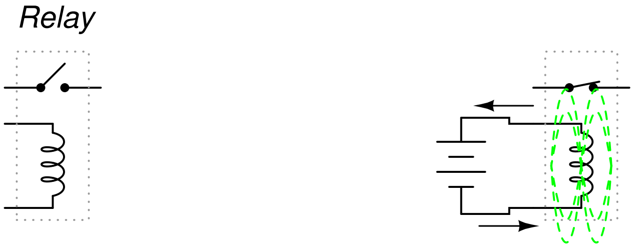
While the magnetic field surrounding a current-carrying wire is indeed interesting, it is quite weak for common amounts of current, able to deflect a compass needle and not much more. To create a stronger magnetic field force (and consequently, more field flux) with the same amount of electric current, we can wrap the wire into a coil shape, where the circling magnetic fields around the wire will join to create a larger field with a definite magnetic (north and south) polarity:



The amount of magnetic field force generated by a coiled wire is proportional to the current through the wire multiplied by the number of "turns" or "wraps" of wire in the coil. This field force is called *magnetomotive force* (mmf), and is very much analogous to electromotive force (E) in an electric circuit.

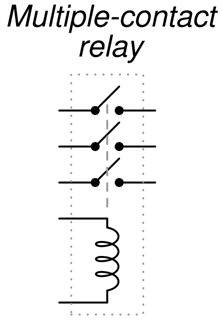
An *electromagnet* is a piece of wire intended to generate a magnetic field with the passage of electric current through it. Though all current-carrying conductors produce magnetic fields, an electromagnet is usually constructed in such a way as to maximize the strength of the magnetic field it produces for a special purpose. Electromagnets find frequent application in research, industry, medical, and consumer products.

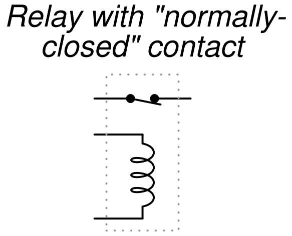
As an electrically-controllable magnet, electromagnets find application in a wide variety of "electromechanical" devices: machines that effect mechanical force or motion through electrical power. Perhaps the most obvious example of such a machine is the *electric motor*.



Applying current through the coil causes the switch to close.

Relays can be constructed to actuate multiple switch contacts, or operate them in "reverse" (energizing the coil will *open* the switch contact, and unpowering the coil will allow it to spring closed again).





15.2 Magnetic units of measurement

If the burden of two systems of measurement for common quantities (English vs. metric) throws your mind into confusion, this is not the place for you! Due to an early lack of standardization in the science of magnetism, we have been plagued with no less than three complete systems of measurement for magnetic quantities.

First, we need to become acquainted with the various quantities associated with magnetism. There are quite a few more quantities to be dealt with in magnetic systems than for electrical systems. With electricity, the basic quantities are Voltage (E), Current (I), Resistance (R), and Power (P). The first three are related to one another by Ohm's Law (E=IR ; I=E/R ; R=E/I), while Power is related to voltage, current, and resistance by Joule's Law (P=IE ; P=I^2R ; P=E^2/R).

With magnetism, we have the following quantities to deal with:

Magnetomotive Force – The quantity of magnetic field force, or ”push.” Analogous to electric voltage (electromotive force).

Field Flux – The quantity of total field effect, or ”substance” of the field. Analogous to electric current.

Field Intensity – The amount of field force (mmf) distributed over the length of the electromagnet. Sometimes referred to as *Magnetizing Force*.

Flux Density – The amount of magnetic field flux concentrated in a given area.

Reluctance – The opposition to magnetic field flux through a given volume of space or material. Analogous to electrical resistance.

Permeability – The specific measure of a material's acceptance of magnetic flux, analogous to the specific resistance of a conductive material (ρ), except inverse (greater permeability means easier passage of magnetic flux, whereas greater specific resistance means more difficult passage of electric current).

But wait . . . the fun is just beginning! Not only do we have more quantities to keep track of with magnetism than with electricity, but we have several different systems of unit measurement for each of these quantities. As with common quantities of length, weight, volume, and temperature, we have both English and metric systems. However, there is actually more than one metric system of units, and multiple metric systems are used in magnetic field measurements! One is called the *cgs*, which stands for **C**entimeter-**G**ram-**S**econd, denoting the root measures upon which the whole system is based. The other was originally known as the *mks* system, which stood for **M**eter-**K**ilogram-**S**econd, which was later revised into another system, called *rmks*, standing for **R**ationalized **M**eter-**K**ilogram-**S**econd. This ended up being adopted as an international standard and renamed *SI* (**S**ysteme **I**nternational).

Quantity	Symbol	Unit of Measurement and abbreviation		
		CGS	SI	English
Field Force	mmf	Gilbert (Gb)	Amp-turn	Amp-turn
Field Flux	Φ	Maxwell (Mx)	Weber (Wb)	Line
Field Intensity	H	Oersted (Oe)	Amp-turns per meter	Amp-turns per inch
Flux Density	B	Gauss (G)	Tesla (T)	Lines per square inch
Reluctance	$\Re$	Gilberts per Maxwell	Amp-turns per Weber	Amp-turns per line
Permeability	μ	Gauss per Oersted	Tesla-meters per Amp-turn	Lines per inch-Amp-turn

And yes, the μ symbol is really the same as the metric prefix ”micro.” I find this especially confusing, using the exact same alphabetical character to symbolize both a specific quantity and a general metric prefix!

As you might have guessed already, the relationship between field force, field flux, and reluctance is much the same as that between the electrical quantities of electromotive force (E), current (I), and resistance (R). This provides something akin to an Ohm's Law for magnetic circuits:

A comparison of "Ohm's Law" for electric and magnetic circuits:

$$E = IR \qquad\qquad mmf = \Phi\Re$$

Electrical Magnetic

And, given that permeability is inversely analogous to specific resistance, the equation for finding the reluctance of a magnetic material is very similar to that for finding the resistance of a conductor:

A comparison of electrical and magnetic opposition:

$$R = \rho\frac{l}{A} \qquad\qquad \Re = \frac{l}{\mu A}$$

Electrical Magnetic

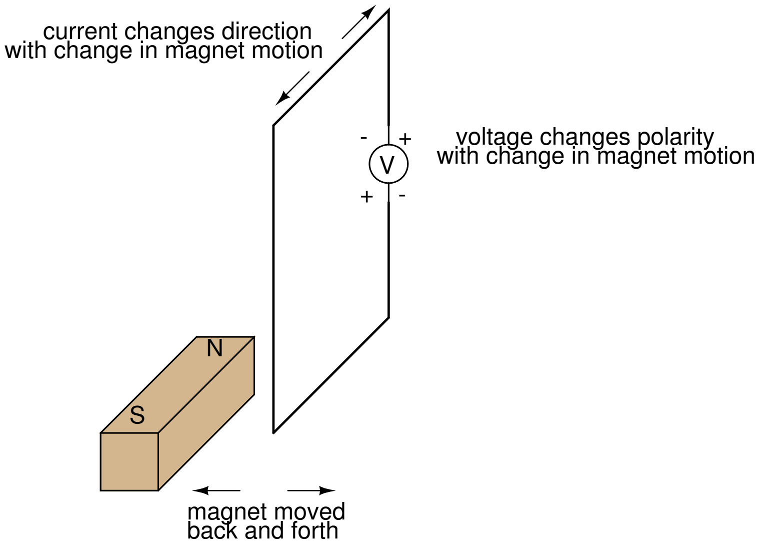
In either case, a longer piece of material provides a greater opposition, all other factors being equal. Also, a larger cross-sectional area makes for less opposition, all other factors being equal.

15.3 Electromagnetic induction

While Oersted's surprising discovery of electromagnetism paved the way for more practical *applications* of electricity, it was Michael Faraday who gave us the key to the practical *generation* of electricity: electromagnetic induction. Faraday discovered that a voltage would be generated across a length of wire if that wire was exposed to a perpendicular magnetic field flux of changing intensity.

An easy way to create a magnetic field of changing intensity is to move a permanent magnet next to a wire or coil of wire. Remember: the magnetic field must increase or decrease in intensity *perpendicular* to the wire (so that the lines of flux "cut across" the conductor), or else no voltage will be induced:

Electromagnetic induction


current changes direction
with change in magnet motion
-
+
V
+
-
voltage changes polarity
with change in magnet motion
N
S
magnet moved
back and forth

Faraday was able to mathematically relate the rate of change of the magnetic field flux with induced voltage (note the use of a lower-case letter "e" for voltage. This refers to *instantaneous* voltage, or voltage at a specific point in time, rather than a steady, stable voltage.):

$$e = N \frac{d\Phi}{dt}$$

Where,

$e =$ (Instantaneous) induced voltage in volts

$N =$ Number of turns in wire coil (straight wire = 1)

$\Phi =$ Magnetic flux in Webers

$t =$ Time in seconds

The "d" terms are standard calculus notation, representing rate-of-change of flux over time.

"N" stands for the number of turns, or wraps, in the wire coil (assuming that the wire is formed in the shape of a coil for maximum electromagnetic efficiency).

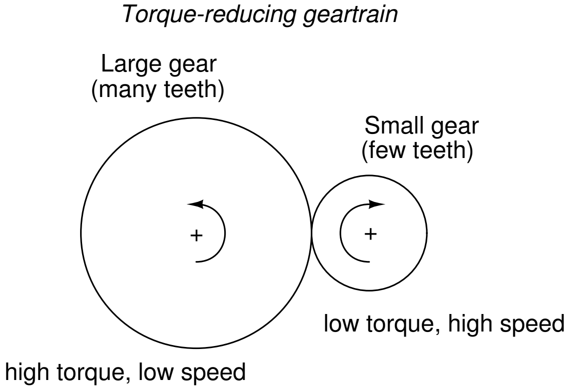
This phenomenon is put into obvious practical use in the construction of electrical generators, which use mechanical power to move a magnetic field past coils of wire to generate voltage. However, this is by no means the only practical use for this principle.

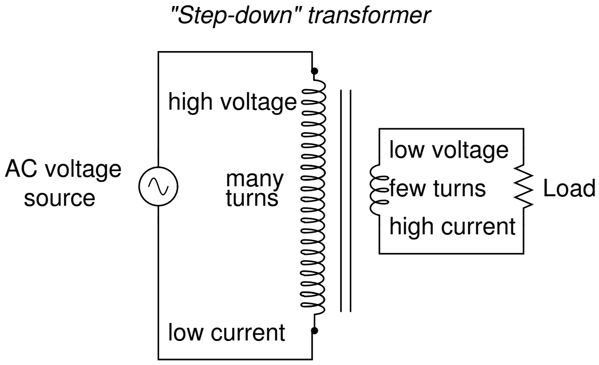
If we recall that the magnetic field produced by a current-carrying wire was always perpendicular to that wire, and that the flux intensity of that magnetic field varied with the amount of current through it, we can see that a wire is capable of inducing a voltage *along its own length* simply due to a change in current through it. This effect is called *self-induction:* a changing magnetic field produced by changes in current through a wire inducing voltage along the length of that same wire. If the magnetic field flux is enhanced by bending the wire into the shape of a coil, and/or wrapping that coil around a material of high permeability, this effect of self-induced voltage will be more intense. A device constructed to take advantage of this effect is called an *inductor*, and will be discussed in greater detail in the next chapter.

A device specifically designed to produce the effect of mutual inductance between two or more coils is called a *transformer.*

Because magnetically-induced voltage only happens when the magnetic field flux is *changing* in strength relative to the wire, mutual inductance between two coils can only happen with alternating (changing – AC) voltage, and not with direct (steady – DC) voltage. The only applications for mutual inductance in a DC system is where some means is available to switch power on and off to the coil (thus creating a *pulsing* DC voltage), the induced voltage peaking at every pulse.

A very useful property of transformers is the ability to transform voltage and current levels according to a simple ratio, determined by the ratio of input and output coil turns. If the energized coil of a transformer is energized by an AC voltage, the amount of AC voltage induced in the unpowered coil will be equal to the input voltage multiplied by the ratio of output to input wire turns in the coils. Conversely, the current through the windings of the output coil compared to the input coil will follow the opposite ratio: if the voltage is increased from input coil to output coil, the current will be decreased by the same proportion. This action of the transformer is analogous to that of mechanical gear, belt sheave, or chain sprocket ratios:



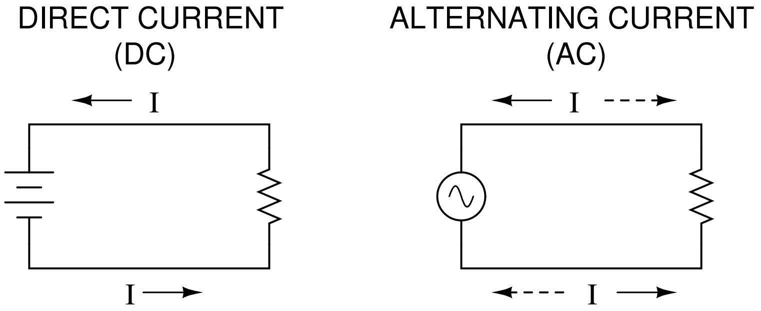


A transformer designed to output more voltage than it takes in across the input coil is called a "step-up" transformer, while one designed to do the opposite is called a "step-down," in reference to the transformation of voltage that takes place. The current through each respective coil, of course, follows the exact opposite proportion.

15.4 AC

Most students of electricity begin their study with what is known as *direct current* (DC), which is electricity flowing in a constant direction, and/or possessing a voltage with constant polarity. DC is the kind of electricity made by a battery (with definite positive and negative terminals), or the kind of charge generated by rubbing certain types of materials against each other.

As useful and as easy to understand as DC is, it is not the only "kind" of electricity in use. Certain sources of electricity (most notably, rotary electro-mechanical generators) naturally produce voltages alternating in polarity, reversing positive and negative over time. Either as a voltage switching polarity or as a current switching direction back and forth, this "kind" of electricity is known as Alternating Current (AC):

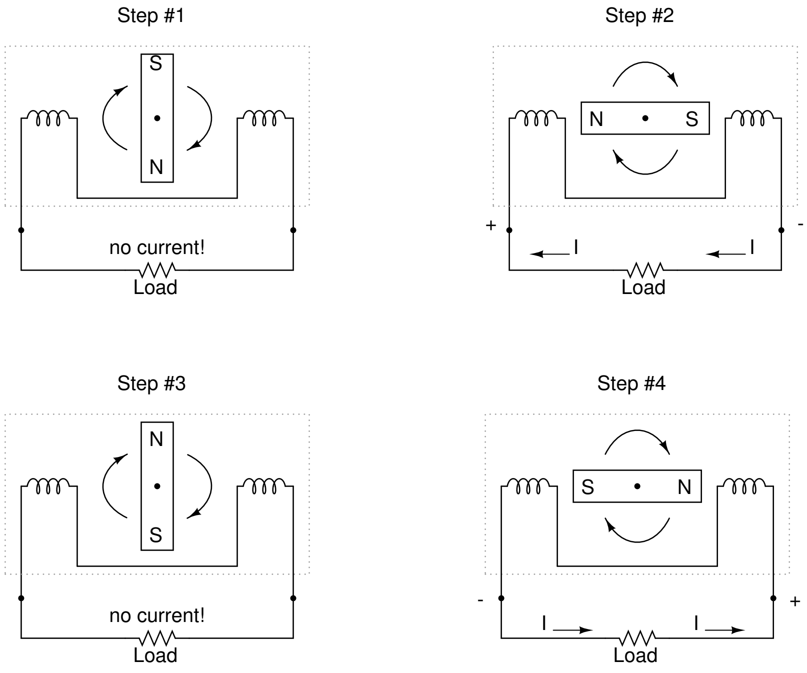


Whereas the familiar battery symbol is used as a generic symbol for any DC voltage source, the circle with the wavy line inside is the generic symbol for any AC voltage source.

One might wonder why anyone would bother with such a thing as AC. It is true that in some cases AC holds no practical advantage over DC. In applications where electricity is used to dissipate energy in the form of heat, the polarity or direction of current is irrelevant, so long as there is enough voltage and current to the load to produce the desired heat (power dissipation). However, with AC it is possible to build electric generators, motors and power distribution systems that are far more efficient than DC, and so we find AC used predominately across the world in high power applications. To explain the details of why this is so, a bit of background knowledge about AC is necessary.

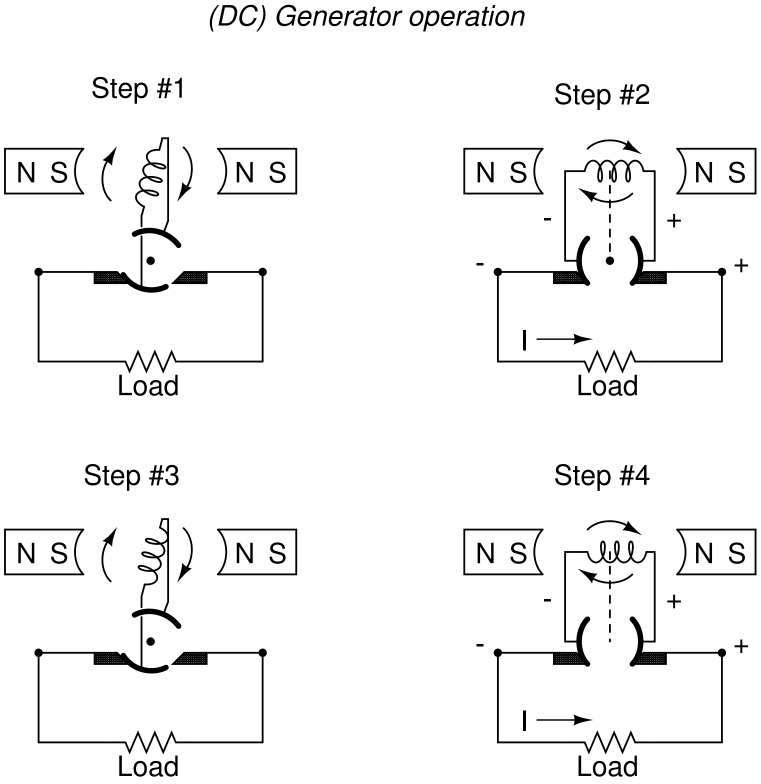
If a machine is constructed to rotate a magnetic field around a set of stationary wire coils with the turning of a shaft, AC voltage will be produced across the wire coils as that shaft is rotated, in accordance with Faraday's Law of electromagnetic induction. This is the basic operating principle of an AC generator, also known as an *alternator*:

Alternator operation



Notice how the polarity of the voltage across the wire coils reverses as the opposite poles of the rotating magnet pass by. Connected to a load, this reversing voltage polarity will create a reversing current direction in the circuit. The faster the alternator's shaft is turned, the faster the magnet will spin, resulting in an alternating voltage and current that switches directions more often in a given amount of time.

While DC generators work on the same general principle of electromagnetic induction, their construction is not as simple as their AC counterparts. With a DC generator, the coil of wire is mounted in the shaft where the magnet is on the AC alternator, and electrical connections are made to this spinning coil via stationary carbon "brushes" contacting copper strips on the rotating shaft. All this is necessary to switch the coil's changing output polarity to the external circuit so the external circuit sees a constant polarity:

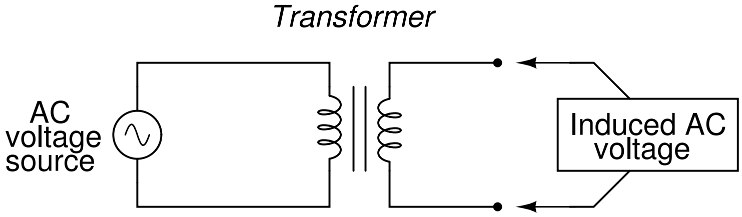


The generator shown above will produce two pulses of voltage per revolution of the shaft, both pulses in the same direction (polarity). In order for a DC generator to produce *constant* voltage, rather than brief pulses of voltage once every 1/2 revolution, there are multiple sets of coils making intermittent contact with the brushes. The diagram shown above is a bit more simplified than what you would see in real life.

The problems involved with making and breaking electrical contact with a moving coil should be obvious (sparking and heat), especially if the shaft of the generator is revolving at high speed. If the atmosphere surrounding the machine contains flammable or explosive vapors, the practical problems of spark-producing brush contacts are even greater. An AC generator (alternator) does not require brushes and commutators to work, and so is immune to these problems experienced by DC generators.

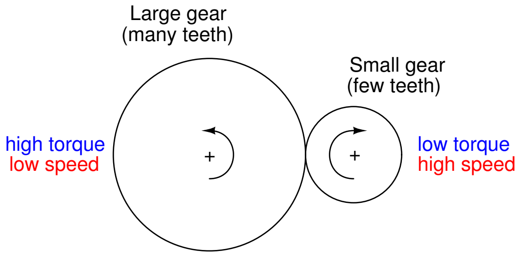
The benefits of AC over DC with regard to generator design is also reflected in electric motors. While DC motors require the use of brushes to make electrical contact with moving coils of wire, AC motors do not. In fact, AC and DC motor designs are very similar to their generator counterparts (identical for the sake of this tutorial), the AC motor being dependent upon the reversing magnetic field produced by alternating current through its stationary coils of wire to rotate the rotating magnet around on its shaft, and the DC motor being dependent on the brush contacts making and breaking connections to reverse current through the rotating coil every 1/2 rotation (180 degrees).

So we know that AC generators and AC motors tend to be simpler than DC generators and DC motors. This relative simplicity translates into greater reliability and lower cost of manufacture. But what else is AC good for? Surely there must be more to it than design details of generators and motors! Indeed there is. There is an effect of electromagnetism known as *mutual induction*, whereby two or more coils of wire placed so that the changing magnetic field created by one induces a voltage in the other. If we have two mutually inductive coils and we energize one coil with AC, we will create an AC voltage in the other coil. When used as such, this device is known as a *transformer*:

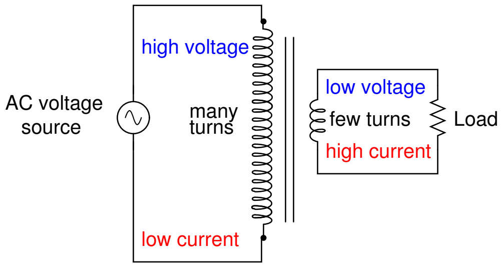


The fundamental significance of a transformer is its ability to step voltage up or down from the powered coil to the unpowered coil. The AC voltage induced in the unpowered (”secondary”) coil is equal to the AC voltage across the powered (”primary”) coil multiplied by the ratio of secondary coil turns to primary coil turns. If the secondary coil is powering a load, the current through the secondary coil is just the opposite: primary coil current multiplied by the ratio of primary to secondary turns. This relationship has a very close mechanical analogy, using torque and speed to represent voltage and current, respectively:

Speed multiplication geartrain

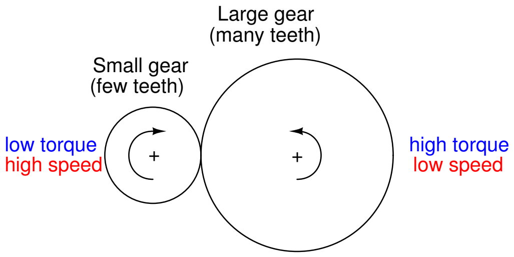


"Step-down" transformer

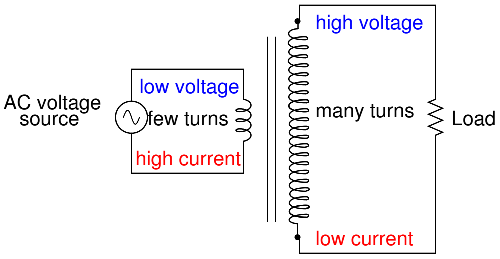


If the winding ratio is reversed so that the primary coil has less turns than the secondary coil, the transformer "steps up" the voltage from the source level to a higher level at the load:

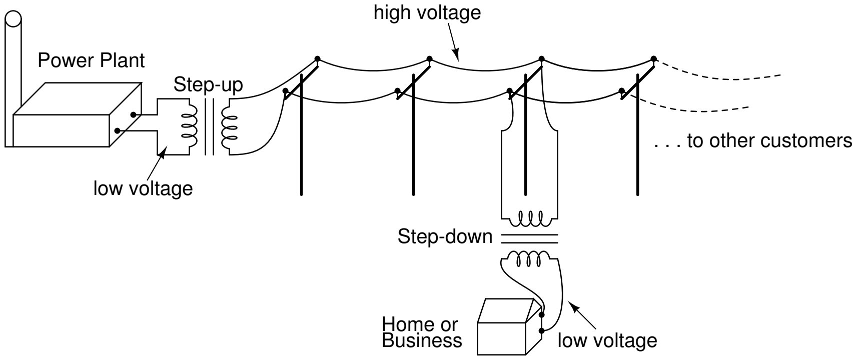
Speed reduction geartrain



"Step-up" transformer



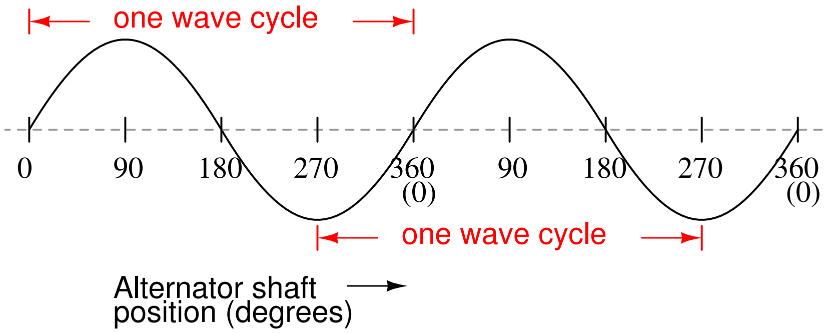
The transformer's ability to step AC voltage up or down with ease gives AC an advantage unmatched by DC in the realm of power distribution. When transmitting electrical power over long distances, it is far more efficient to do so with stepped-up voltages and stepped-down currents (smaller-diameter wire with less resistive power losses), then step the voltage back down and the current back up for industry, business, or consumer use use.



Transformer technology has made long-range electric power distribution practical. Without the ability to efficiently step voltage up and down, it would be cost-prohibitive to construct power systems for anything but close-range (within a few miles at most) use.

As useful as transformers are, they only work with AC, not DC. Because the phenomenon of mutual inductance relies on *changing* magnetic fields, and direct current (DC) can only produce steady magnetic fields, transformers simply will not work with direct current. Of course, direct current may be interrupted (pulsed) through the primary winding of a transformer to create a changing magnetic field (as is done in automotive ignition systems to produce high-voltage spark plug power from a low-voltage DC battery), but pulsed DC is not that different from AC. Perhaps more than any other reason, this is why AC finds such widespread application in power systems.

If we were to follow the changing voltage produced by a coil in an alternator from any point on the sine wave graph to that point when the wave shape begins to repeat itself, we would have marked exactly one *cycle* of that wave. This is most easily shown by spanning the distance between identical peaks, but may be measured between any corresponding points on the graph. The degree marks on the horizontal axis of the graph represent the domain of the trigonometric sine function, and also the angular position of our simple two-pole alternator shaft as it rotates:



Since the horizontal axis of this graph can mark the passage of time as well as shaft position in degrees, the dimension marked for one cycle is often measured in a unit of time, most often seconds or fractions of a second. When expressed as a measurement, this is often called the *period* of a wave. The period of a wave in degrees is *always* 360, but the amount of time one period occupies depends on the rate voltage oscillates back and forth.

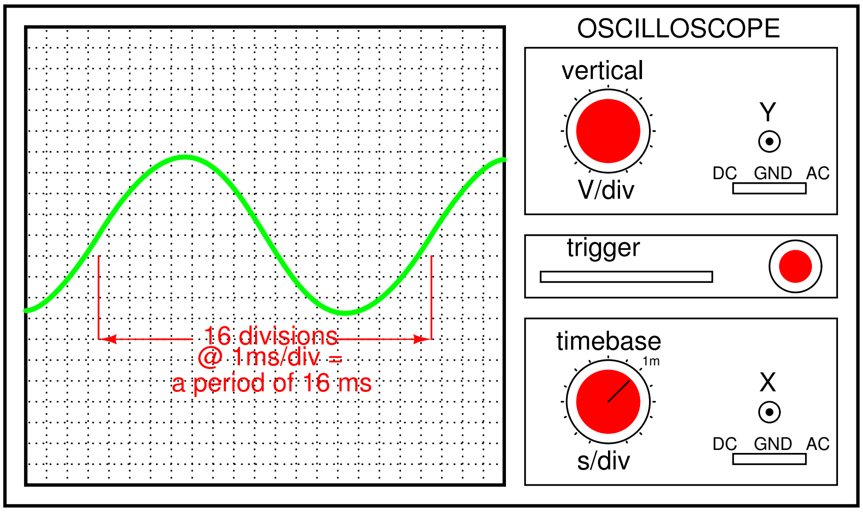
A more popular measure for describing the alternating rate of an AC voltage or current wave than *period* is the rate of that back-and-forth oscillation. This is called *frequency*. The modern unit for frequency is the Hertz (abbreviated Hz), which represents the number of wave cycles completed during one second of time. In the United States of America, the standard power-line frequency is 60 Hz, meaning that the AC voltage oscillates at a rate of 60 complete back-and-forth cycles every second. In Europe, where the power system frequency is 50 Hz, the AC voltage only completes 50 cycles every second. A radio station transmitter broadcasting at a frequency of 100 MHz generates an AC voltage oscillating at a rate of 100 *million* cycles every second.

Prior to the canonization of the Hertz unit, frequency was simply expressed as "cycles per second." Older meters and electronic equipment often bore frequency units of "CPS" (Cycles Per Second) instead of Hz. Many people believe the change from self-explanatory units like CPS to Hertz constitutes a step backward in clarity. A similar change occurred when the unit of "Celsius" replaced that of "Centigrade" for metric temperature measurement. The name Centigrade was based on a 100-count ("Centi-") scale ("-grade") representing the melting and boiling points of H_2O, respectively. The name Celsius, on the other hand, gives no hint as to the unit's origin or meaning.

Period and frequency are mathematical reciprocals of one another. That is to say, if a wave has a period of 10 seconds, its frequency will be 0.1 Hz, or 1/10 of a cycle per second:

$$\text{Frequency in Hertz} = \frac{1}{\text{Period in seconds}}$$

An instrument called an *oscilloscope* is used to display a changing voltage over time on a graphical screen. You may be familiar with the appearance of an *ECG* or *EKG* (electrocardiograph) machine, used by physicians to graph the oscillations of a patient's heart over time. The ECG is a special-purpose oscilloscope expressly designed for medical use. General-purpose oscilloscopes have the ability to display voltage from virtually any voltage source, plotted as a graph with time as the independent variable. The relationship between period and frequency is very useful to know when displaying an AC voltage or current waveform on an oscilloscope screen. By measuring the period of the wave on the horizontal axis of the oscilloscope screen and reciprocating that time value (in seconds), you can determine the frequency in Hertz.



$$\text{Frequency} = \frac{1}{\text{period}} = \frac{1}{16\text{ ms}} = 62.5\text{ Hz}$$

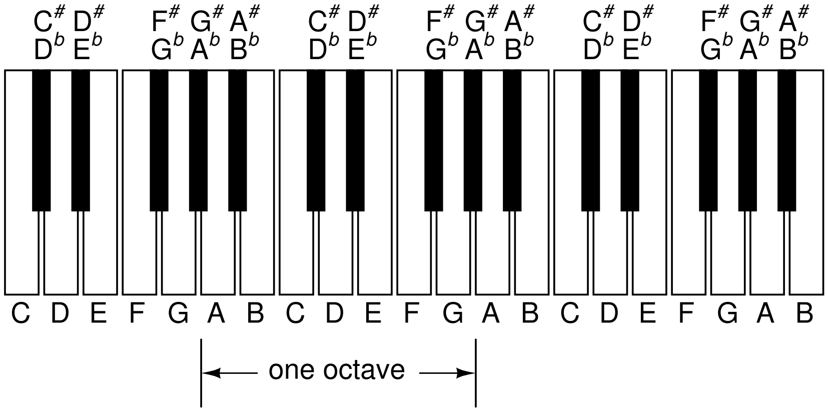
Voltage and current are by no means the only physical variables subject to variation over time. Much more common to our everyday experience is *sound*, which is nothing more than the alternating compression and decompression (pressure waves) of air molecules, interpreted by our ears as a physical sensation. Because alternating current is a wave phenomenon, it shares many of the properties of other wave phenomena, like sound. For this reason, sound (especially structured music) provides an excellent analogy for relating AC concepts.

In musical terms, frequency is equivalent to *pitch*. Low-pitch notes such as those produced by a tuba or bassoon consist of air molecule vibrations that are relatively slow (low frequency). High-pitch notes such as those produced by a flute or whistle consist of the same type of vibrations in the air, only vibrating at a much faster rate (higher frequency). Here is a table showing the actual frequencies for a range of common musical notes:

Note	Musical designation	Frequency (in hertz)
A	A_1	220.00
A sharp (or B flat)	$A^{\#}$ or B^{b}	233.08
B	B_1	246.94
C (middle)	C	261.63
C sharp (or D flat)	$C^{\#}$ or D^{b}	277.18
D	D	293.66
D sharp (or E flat)	$D^{\#}$ or E^{b}	311.13
E	E	329.63
F	F	349.23
F sharp (or G flat)	$F^{\#}$ or G^{b}	369.99
G	G	392.00
G sharp (or A flat)	$G^{\#}$ or A^{b}	415.30
A	A	440.00
A sharp (or B flat)	$A^{\#}$ or B^{b}	466.16
B	B	493.88
C	C^1	523.25

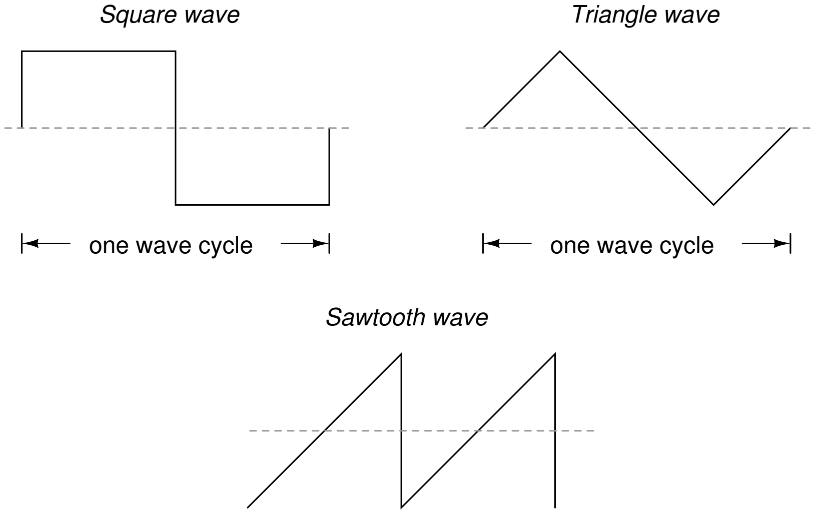
Astute observers will notice that all notes on the table bearing the same letter designation are related by a frequency ratio of 2:1. For example, the first frequency shown (designated with the letter "A") is 220 Hz. The next highest "A" note has a frequency of 440 Hz – exactly twice as many sound wave cycles per second. The same 2:1 ratio holds true for the first A sharp (233.08 Hz) and the next A sharp (466.16 Hz), and for all note pairs found in the table.

Audibly, two notes whose frequencies are exactly double each other sound remarkably similar. This similarity in sound is musically recognized, the shortest span on a musical scale separating such note pairs being called an *octave*. Following this rule, the next highest "A" note (one octave above 440 Hz) will be 880 Hz, the next lowest "A" (one octave below 220 Hz) will be 110 Hz. A view of a piano keyboard helps to put this scale into perspective:


C# D# F# G# A# C# D# F# G# A# C# D# F# G# A#
Db Eb Gb Ab Bb Db Eb Gb Ab Bb Db Eb Gb Ab Bb
C D E F G A B C D E F G A B C D E F G A B
one octave

As you can see, one octave is equal to *eight* white keys' worth of distance on a piano keyboard. The familiar musical mnemonic (doe-ray-mee-fah-so-lah-tee-doe) – yes, the same pattern immortalized in the whimsical Rodgers and Hammerstein song sung in The Sound of Music – covers one octave from C to C.

While electromechanical alternators and many other physical phenomena naturally produce sine waves, this is not the only kind of alternating wave in existence. Other "waveforms" of AC are commonly produced within electronic circuitry. Here are but a few sample waveforms and their common designations:



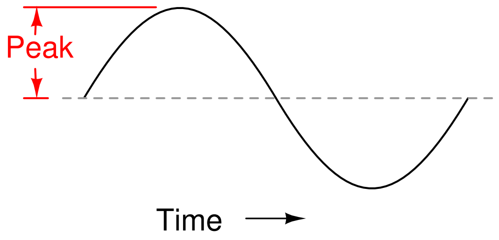
These waveforms are by no means the only kinds of waveforms in existence. They're simply a few that are common enough to have been given distinct names. Even in circuits that are supposed to manifest "pure" sine, square, triangle, or sawtooth voltage/current waveforms, the real-life result is often a distorted version of the intended waveshape. Some waveforms are so complex that they defy classification as a particular "type" (including waveforms associated with many kinds of musical instruments). Generally speaking, any waveshape bearing close resemblance to a perfect sine wave is termed *sinusoidal*, anything different being labeled as *non-sinusoidal*. Being that the waveform of an AC voltage or current is crucial to its impact in a circuit, we need to be aware of the fact that AC waves come in a variety of shapes.

15.5 Measurements of AC magnitude

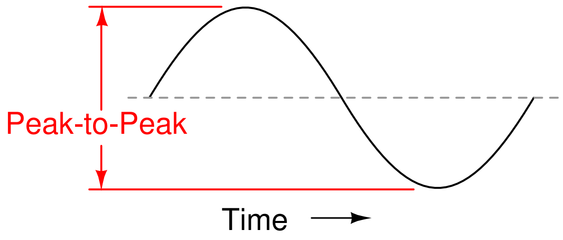
So far we know that AC voltage alternates in polarity and AC current alternates in direction. We also know that AC can alternate in a variety of different ways, and by tracing the alternation over time we can plot it as a "waveform." We can measure the rate of alternation by measuring the time it takes for a wave to evolve before it repeats itself (the "period"), and express this as cycles per unit time, or "frequency." In music, frequency is the same as *pitch*, which is the essential property distinguishing one note from another.

However, we encounter a measurement problem if we try to express how large or small an AC quantity is. With DC, where quantities of voltage and current are generally stable, we have little trouble expressing how much voltage or current we have in any part of a circuit. But how do you grant a single measurement of magnitude to something that is constantly changing?

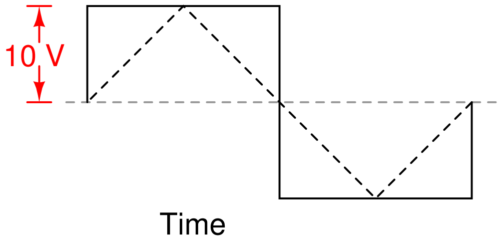
One way to express the intensity, or magnitude (also called the *amplitude*), of an AC quantity is to measure its peak height on a waveform graph. This is known as the *peak* or *crest* value of an AC waveform:

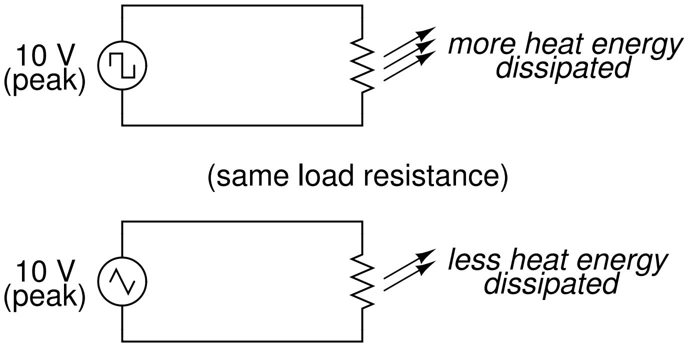


Another way is to measure the total height between opposite peaks. This is known as the *peak-to-peak* (P-P) value of an AC waveform:

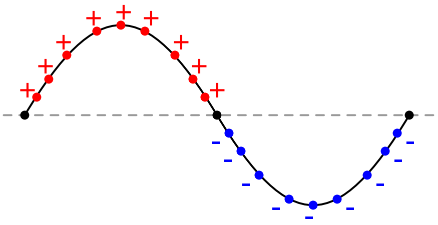


Unfortunately, either one of these expressions of waveform amplitude can be misleading when comparing two different types of waves. For example, a square wave peaking at 10 volts is obviously a greater amount of voltage for a greater amount of time than a triangle wave peaking at 10 volts. The effects of these two AC voltages powering a load would be quite different:



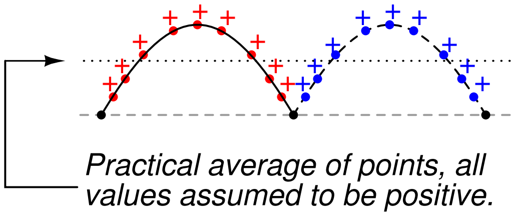


One way of expressing the amplitude of different waveshapes in a more equivalent fashion is to mathematically average the values of all the points on a waveform's graph to a single, aggregate number. This amplitude measure is known simply as the *average* value of the waveform. If we average all the points on the waveform algebraically (that is, to consider their *sign*, either positive or negative), the average value for most waveforms is technically zero, because all the positive points cancel out all the negative points over a full cycle:



*True average value of all points (considering their signs) is **zero**!*

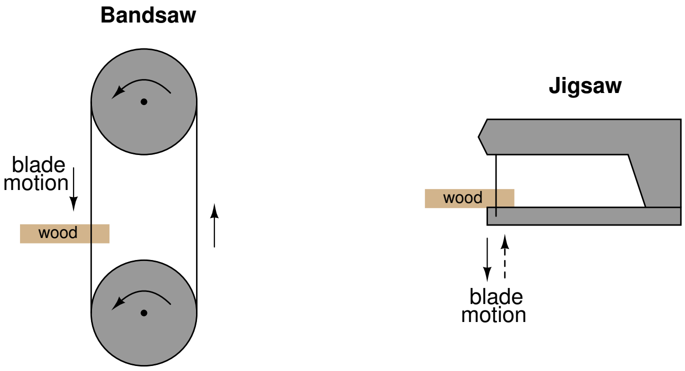
This, of course, will be true for any waveform having equal-area portions above and below the "zero" line of a plot. However, as a *practical* measure of a waveform's aggregate value, "average" is usually defined as the mathematical mean of all the points' *absolute values* over a cycle. In other words, we calculate the practical average value of the waveform by considering all points on the wave as positive quantities, as if the waveform looked like this:


Practical average of points, all values assumed to be positive.

Polarity-insensitive mechanical meter movements (meters designed to respond equally to the positive and negative half-cycles of an alternating voltage or current) register in proportion to the waveform's (practical) average value, because the inertia of the pointer against the tension of the spring naturally averages the force produced by the varying voltage/current values over time. Conversely, polarity-sensitive meter movements vibrate uselessly if exposed to AC voltage or current, their needles oscillating rapidly about the zero mark, indicating the true (algebraic) average value of zero for a symmetrical waveform. When the "average" value of a waveform is referenced in this text, it will be assumed that the "practical" definition of average is intended unless otherwise specified.

Another method of deriving an aggregate value for waveform amplitude is based on the waveform's ability to do useful work when applied to a load resistance. Unfortunately, an AC measurement based on work performed by a waveform is not the same as that waveform's "average" value, because the *power* dissipated by a given load (work performed per unit time) is not directly proportional to the magnitude of either the voltage or current impressed upon it. Rather, power is proportional to the *square* of the voltage or current applied to a resistance (P = E^2/R, and P = I^2R). Although the mathematics of such an amplitude measurement might not be straightforward, the utility of it is.

Consider a bandsaw and a jigsaw, two pieces of modern woodworking equipment. Both types of saws cut with a thin, toothed, motor-powered metal blade to cut wood. But while the bandsaw uses a continuous motion of the blade to cut, the jigsaw uses a back-and-forth motion. The comparison of alternating current (AC) to direct current (DC) may be likened to the comparison of these two saw types:

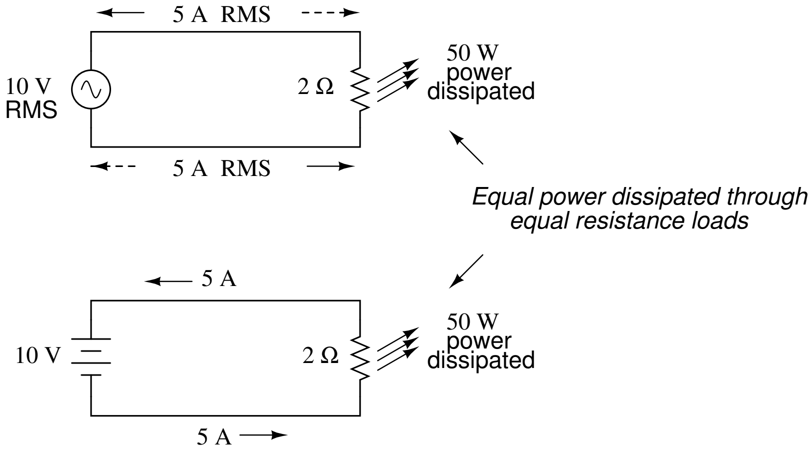


(analogous to DC)

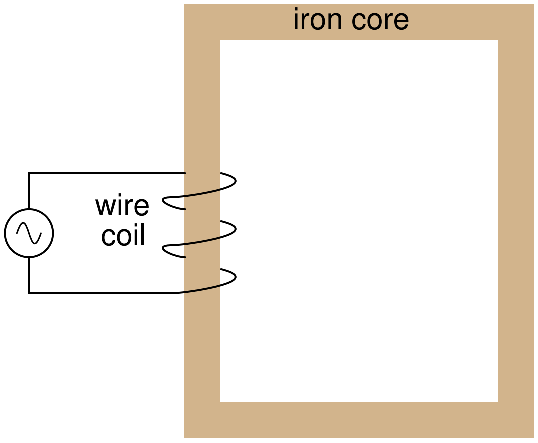
(analogous to AC)

The problem of trying to describe the changing quantities of AC voltage or current in a single, aggregate measurement is also present in this saw analogy: how might we express the speed of a jigsaw blade? A bandsaw blade moves with a constant speed, similar to the way DC voltage pushes or DC current moves with a constant magnitude. A jigsaw blade, on the other hand, moves back and forth, its blade speed constantly changing. What is more, the back-and-forth motion of any two jigsaws may not be of the same type, depending on the mechanical design of the saws. One jigsaw might move its blade with a sine-wave motion, while another with a triangle-wave motion. To rate a jigsaw based on its *peak* blade speed would be quite misleading when comparing one jigsaw to another (or a jigsaw with a bandsaw!). Despite the fact that these different saws move their blades in different manners, they are equal in one respect: they all cut wood, and a quantitative comparison of this common function can serve as a common basis for which to rate blade speed.

Picture a jigsaw and bandsaw side-by-side, equipped with identical blades (same tooth pitch, angle, etc.), equally capable of cutting the same thickness of the same type of wood at the same rate. We might say that the two saws were equivalent or equal in their cutting capacity. Might this comparison be used to assign a ”bandsaw equivalent” blade speed to the jigsaw's back-and-forth blade motion; to relate the wood-cutting effectiveness of one to the other? This is the general idea used to assign a ”DC equivalent” measurement to any AC voltage or current: whatever magnitude of DC voltage or current would produce the same amount of heat energy dissipation through an equal resistance:



Suppose we were to wrap a coil of insulated wire around a loop of ferromagnetic material and energize this coil with an AC voltage source:



As an inductor, we would expect this iron-core coil to oppose the applied voltage with its inductive reactance, limiting current through the coil as predicted by the equations $X_L = 2\pi fL$ and I=E/X (or I=E/Z). For the purposes of this example, though, we need to take a more detailed look at the interactions of voltage, current, and magnetic flux in the device.

Kirchhoff's voltage law describes how the algebraic sum of all voltages in a loop must equal zero. In this example, we could apply this fundamental law of electricity to describe the respective voltages of the source and of the inductor coil. Here, as in any one-source, one-load circuit, the voltage dropped across the load must equal the voltage supplied by the source, assuming zero voltage dropped along the resistance of any connecting wires. In other words, the load (inductor coil) must produce an opposing voltage equal in magnitude to the source, in order that it may balance against the source voltage and produce an algebraic loop voltage sum of zero. From where does this opposing voltage arise? If the load were a resistor, the opposing voltage would originate from the "friction" of electrons flowing through the resistance of the resistor. With a perfect inductor (no resistance in the coil wire), the opposing voltage comes from another mechanism: the *reaction* to a changing magnetic flux in the iron core.

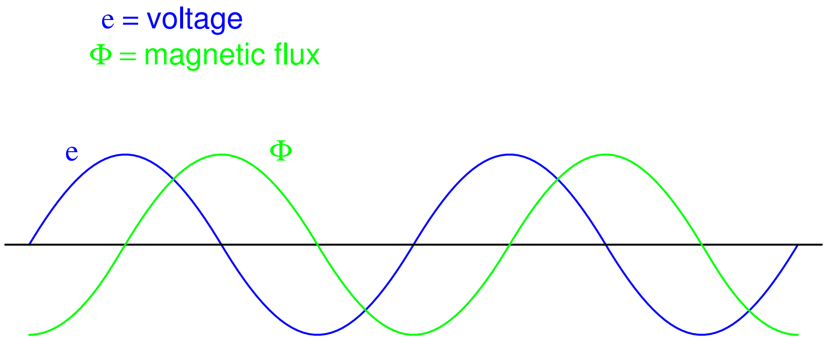
Michael Faraday discovered the mathematical relationship between magnetic flux (Φ) and induced voltage with this equation:

$$e = N \frac{d\Phi}{dt}$$

Where,

$e =$ (Instantaneous) induced voltage in volts
$N =$ Number of turns in wire coil (straight wire = 1)
$\Phi =$ Magnetic flux in Webers
$t =$ Time in seconds

The instantaneous voltage (voltage dropped at any instant in time) across a wire coil is equal to the number of turns of that coil around the core (N) multiplied by the instantaneous rate-of-change in magnetic flux (dΦ/dt) linking with the coil. Graphed, this shows itself as a set of sine waves (assuming a sinusoidal voltage source), the flux wave 90^o lagging behind the voltage wave:



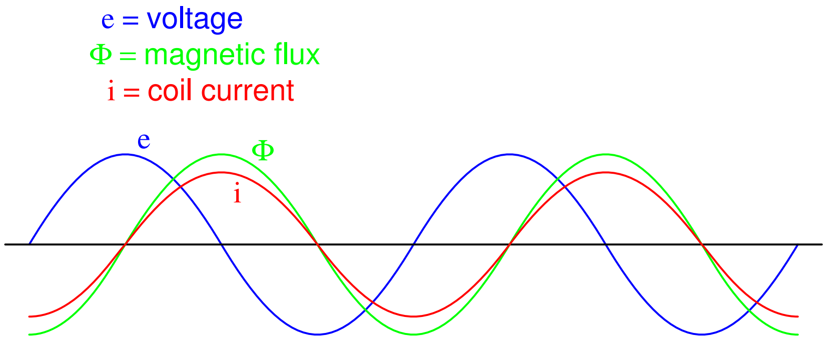
Magnetic flux through a ferromagnetic material is analogous to current through a conductor: it must be motivated by some force in order to occur. In electric circuits, this motivating force is voltage (a.k.a. electromotive force, or EMF). In magnetic "circuits," this motivating force is *magnetomotive force*, or *mmf*. Magnetomotive force (mmf) and magnetic flux (Φ) are related to each other by a property of magnetic materials known as *reluctance* (the latter quantity symbolized by a strange-looking letter "R"):

A comparison of "Ohm's Law" for electric and magnetic circuits:

$E = IR$ — Electrical

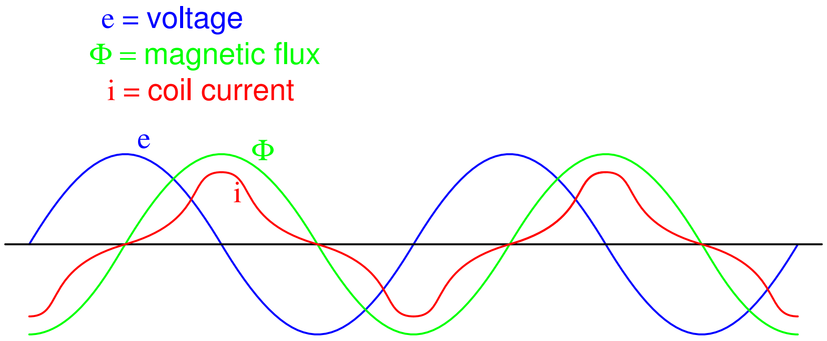
$mmf = \Phi\mathfrak{R}$ — Magnetic

In our example, the mmf required to produce this changing magnetic flux (Φ) must be supplied by a changing current through the coil. Magnetomotive force generated by an electromagnet coil is equal to the amount of current through that coil (in amps) multiplied by the number of turns of that coil around the core (the SI unit for mmf is the *amp-turn*). Because the mathematical relationship between magnetic flux and mmf is directly proportional, and because the mathematical relationship between mmf and current is also directly proportional (no rates-of-change present in either equation), the current through the coil will be in-phase with the flux wave:



This is why alternating current through an inductor lags the applied voltage waveform by 90^o: because that is what is required to produce a changing magnetic flux whose rate-of-change produces an opposing voltage in-phase with the applied voltage. Due to its function in providing magnetizing force (mmf) for the core, this current is sometimes referred to as the *magnetizing current*.

It should be mentioned that the current through an iron-core inductor is not perfectly sinusoidal (sine-wave shaped), due to the nonlinear B/H magnetization curve of iron. In fact, if the inductor is cheaply built, using as little iron as possible, the magnetic flux density might reach high levels (approaching saturation), resulting in a magnetizing current waveform that looks something like this:

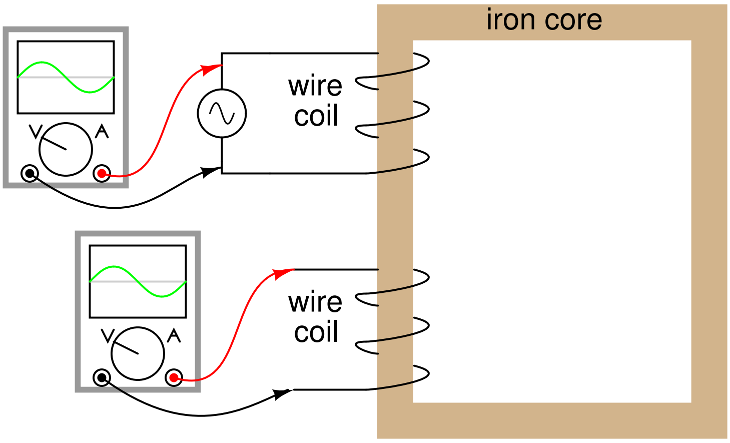


When a ferromagnetic material approaches magnetic flux saturation, disproportionately greater levels of magnetic field force (mmf) are required to deliver equal increases in magnetic field flux (Φ). Because mmf is proportional to current through the magnetizing coil (mmf = NI, where "N" is the number of turns of wire in the coil and "I" is the current through it), the large increases of mmf required to supply the needed increases in flux results in large increases in coil current. Thus, coil current increases dramatically at the peaks in order to maintain a flux waveform that isn't distorted, accounting for the bell-shaped half-cycles of the current waveform in the above plot.

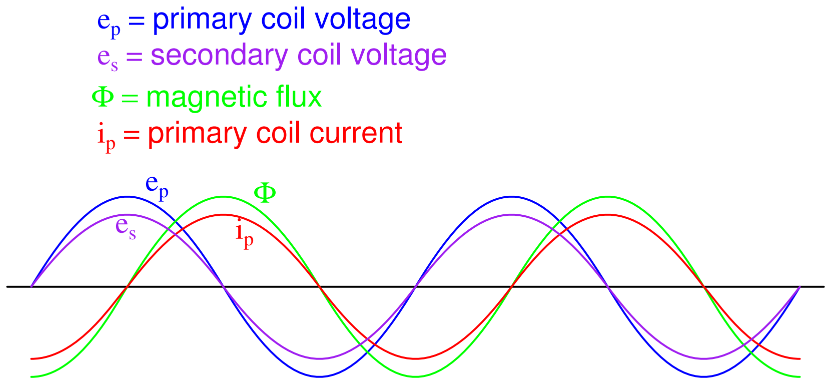
The situation is further complicated by energy losses within the iron core. The effects of hysteresis and eddy currents conspire to further distort and complicate the current waveform, making it even less sinusoidal and altering its phase to be lagging slightly less than 90^o behind the applied voltage waveform. This coil current resulting from the sum total of all magnetic effects in the core ($d\Phi/dt$ magnetization plus hysteresis losses, eddy current losses, etc.) is called the *exciting current*. The distortion of an iron-core inductor's exciting current may be minimized if it is designed for and operated at very low flux densities. Generally speaking, this requires a core with large cross-sectional area, which tends to make the inductor bulky and expensive. For the sake of simplicity, though, we'll assume that our example core is far from saturation and free from all losses, resulting in a perfectly sinusoidal exciting current.

As we've seen already in the inductors chapter, having a current waveform 90^o out of phase with the voltage waveform creates a condition where power is alternately absorbed and returned to the circuit by the inductor. If the inductor is perfect (no wire resistance, no magnetic core losses, etc.), it will dissipate zero power.

Let us now consider the same inductor device, except this time with a second coil wrapped around the same iron core. The first coil will be labeled the *primary* coil, while the second will be labeled the *secondary*:



If this secondary coil experiences the same magnetic flux change as the primary (which it should, assuming perfect containment of the magnetic flux through the common core), and has the same number of turns around the core, a voltage of equal magnitude and phase to the applied voltage will be induced along its length. In the following graph, the induced voltage waveform is drawn slightly smaller than the source voltage waveform simply to distinguish one from the other:


e_p = primary coil voltage
e_s = secondary coil voltage
Φ = magnetic flux
i_p = primary coil current
e_p
Φ
e_s
i_p

Chapter 16

Electronics

(NOTE TO SELF: Mark: I have very little idea of how to make this flow, fit in or even how best to explain any of it. All the content in here is just trawled from other GFDL projects: www.wikipedia.com www.wikibooks.com The syllabus document has NO meaningful information on this stuff.)

Electronics:

16.1 capacitive and inductive circuits

16.1.1 A capacitor

(NOTE TO SELF: If we are going to talk of capacitive circuits we need a definition of capacitor.)

A capacitor (historically known as a "condenser") is a device that stores energy in an electric field, by accumulating an internal imbalance of electric charge. It is made of two conductors separated by a dielectric (insulator). The problem of two parallel plates with a uniform electric field between them is a capacitor.

When voltage exists one end of the capacitor is getting drained and the other end is getting filled with charge. This is known as charging. Charging creates a charge imbalance between the two plates and creates a reverse voltage that stops the capacitor from charging. This is why when capacitors are first connected to voltage charge flows only to stop as the capacitor becomes charged. When a capacitor is charged current stops flowing and it becomes an open circuit. It is as if the capacitor gained infinite resistance.

Just as the capacitor charges it can be discharged.

16.1.2 An inductor

(NOTE TO SELF: If we are going to talk of inductive circuits we need a definition of a inductor)

An inductor is a device which stores energy in a magnetic field. Inductors are formed of a coil of conductive material. When current flows through the wire it creates a magnetic field which exists inside the coil. When the current stops the magnetic field gets less, but we have learnt that a changing magnetic field induces a current in a wire. So when the current turns off the magnetic field decreases inducing another current in the wire. As the field decreases in strength so does the induced magnetic field.

Normally they are made of copper wire, but not always (Example: aluminum wire, or spiral pattern etched on circuit board). The material around and within the coil affects its properties;

common types are air-core (only a coil of wire), iron-core, and ferrite core. Iron and ferrite types are more efficient because they conduct the magnetic field much better than air; of the two, ferrite is more efficient because stray electricity cannot flow through it.

Interesting Fact: Some inductors have more than a core, which is just a rod the coil is formed about. Some are formed like transformers, using two E-shaped pieces facing each other, the wires wound about the central leg of the E's. The E's are made of laminated iron/steel or ferrite.

Important qualities of an inductor

There are several important properties for an inductor.

* Current carrying capacity is determined by wire thickness. * Q, or quality, is determined by the uniformity of the windings, as well as the core material and how thoroughly it surrounds the coil. * Last but not least, the inductance of the coil.

The inductance is determined by several factors.

* coil shape: short and squat is best * core material * windings: winding in opposite directions will cancel out the inductance effect, and you will have only a resistor.

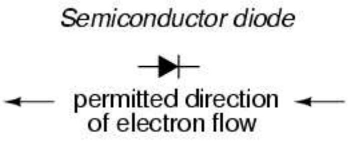
16.2 filters and signal tuning

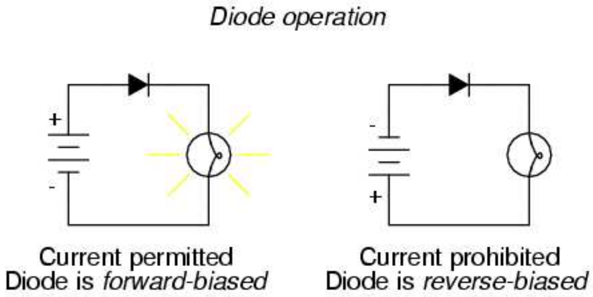
(NOTE TO SELF: I think this relies on an understanding of second order ODEs and thats beyond the scope of the maths syllabus - we can put something high level but there is no way they'll understand it properly - surely we should teach as little phenomonology as possible - the waves chapter has a ton of it already)

16.3 active circuit elements, diode, LED and field effect transistor, operational amplifier

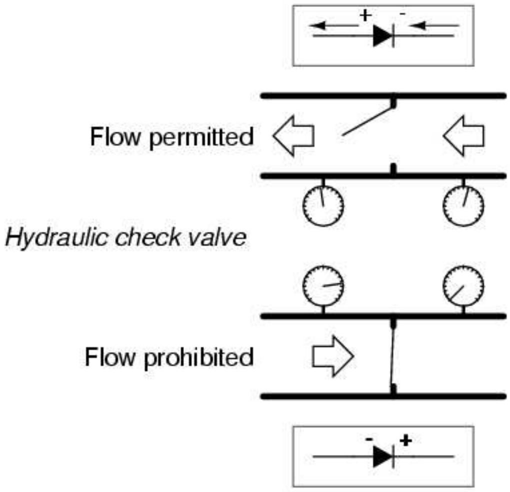
16.3.1 Diode

A diode functions as the electronic version of a one-way valve. By restricting the direction of movement of charge carriers, it allows an electric current to flow in one direction, but blocks it in the opposite direction. It is a one-way street for current.



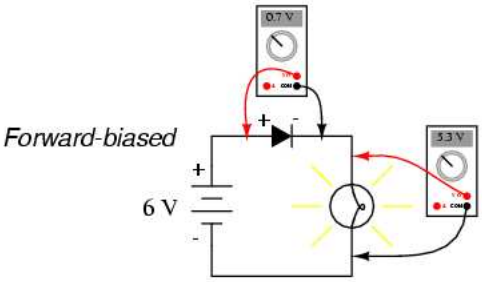


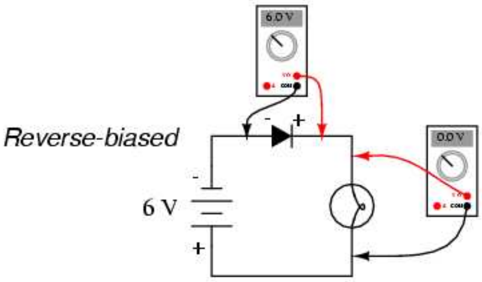
Diode behavior is analogous to the behavior of a hydraulic device called a check valve. A check valve allows fluid flow through it in one direction only:



Check valves are essentially pressure-operated devices: they open and allow flow if the pressure across them is of the correct "polarity" to open the gate (in the analogy shown, greater fluid pressure on the right than on the left). If the pressure is of the opposite "polarity," the pressure difference across the check valve will close and hold the gate so that no flow occurs.

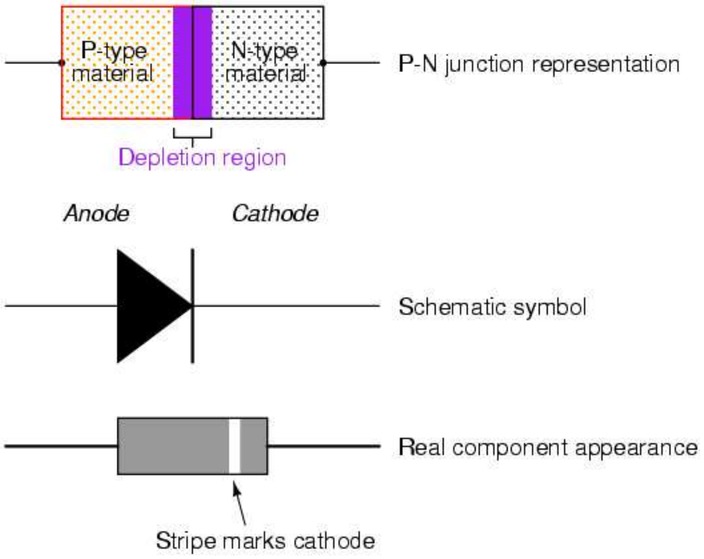
Like check valves, diodes are essentially "pressure-" operated (voltage-operated) devices. The essential difference between forward-bias and reverse-bias is the polarity of the voltage dropped across the diode. Let's take a closer look at the simple battery-diode-lamp circuit shown earlier, this time investigating voltage drops across the various components:





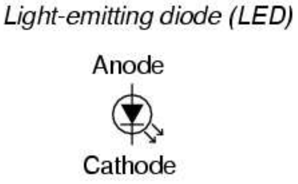
When the diode is forward-biased and conducting current, there is a small voltage dropped across it, leaving most of the battery voltage dropped across the lamp. When the battery's polarity is reversed and the diode becomes reverse-biased, it drops all of the battery's voltage and leaves none for the lamp. If we consider the diode to be a sort of self-actuating switch (closed in the forward-bias mode and open in the reverse-bias mode), this behavior makes sense. The most substantial difference here is that the diode drops a lot more voltage when conducting than the average mechanical switch (0.7 volts versus tens of millivolts).

This forward-bias voltage drop exhibited by the diode is due to the action of the depletion region formed by the P-N junction under the influence of an applied voltage. When there is no voltage applied across a semiconductor diode, a thin depletion region exists around the region of the P-N junction, preventing current through it. The depletion region is for the most part devoid of available charge carriers and so acts as an insulator:



16.3.2 LED

A light-emitting diode (LED) is a semiconductor 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.



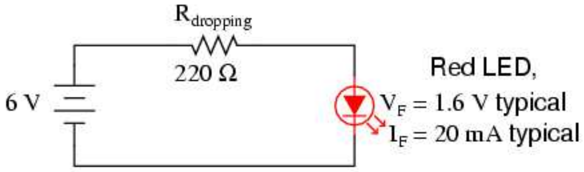
This notation of having two small arrows pointing away from the device is common to the schematic symbols of all light-emitting semiconductor devices. Conversely, if a device is light-activated (meaning that incoming light stimulates it), then the symbol will have two small arrows pointing toward it. 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.

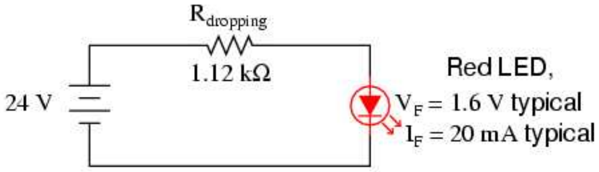
Interesting Fact: Nick Holonyak Jr. (1928) of the University of Illinois at Urbana-Champaign developed the first practical visible-spectrum LED in 1962.

Physical function

Because LEDs are made of different chemical substances than normal rectifying diodes, their forward voltage drops will be different. Typically, LEDs have much larger forward voltage drops than rectifying diodes, anywhere from about 1.6 volts to over 3 volts, depending on the color. Typical operating current for a standard-sized LED is around 20 mA. When operating an LED from a DC voltage source greater than the LED's forward voltage, a series-connected "dropping" resistor must be included to prevent full source voltage from damaging the LED. Consider this example circuit:



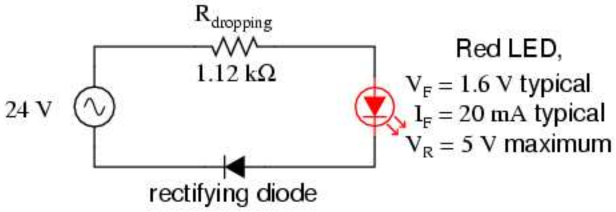
With the LED dropping 1.6 volts, there will be 4.4 volts dropped across the resistor. Sizing the resistor for an LED current of 20 mA is as simple as taking its voltage drop (4.4 volts) and dividing by circuit current (20 mA), in accordance with Ohm's Law (R=E/I). This gives us a figure of 220 ?. Calculating power dissipation for this resistor, we take its voltage drop and multiply by its current (P=IE), and end up with 88 mW, well within the rating of a 1/8 watt resistor. Higher battery voltages will require larger-value dropping resistors, and possibly higher-power rating resistors as well. Consider this example for a supply voltage of 24 volts:



Here, the dropping resistor must be increased to a size of 1.12 k? in order to drop 22.4 volts at 20 mA so that the LED still receives only 1.6 volts. This also makes for a higher resistor power dissipation: 448 mW, nearly one-half a watt of power! Obviously, a resistor rated for 1/8 watt power dissipation or even 1/4 watt dissipation will overheat if used here.

Dropping resistor values need not be precise for LED circuits. Suppose we were to use a 1 k? resistor instead of a 1.12 k? resistor in the circuit shown above. The result would be a slightly greater circuit current and LED voltage drop, resulting in a brighter light from the LED and slightly reduced service life. A dropping resistor with too much resistance (say, 1.5 k? instead of 1.12 k?) will result in less circuit current, less LED voltage, and a dimmer light. LEDs are quite tolerant of variation in applied power, so you need not strive for perfection in sizing the dropping resistor.

Also because of their unique chemical makeup, LEDs have much, much lower peak-inverse voltage (PIV) ratings than ordinary rectifying diodes. A typical LED might only be rated at 5 volts in reverse-bias mode. Therefore, when using alternating current to power an LED, you should connect a protective rectifying diode in series with the LED to prevent reverse breakdown every other half-cycle:



Light emission

The wavelength of the light emitted, and therefore its color, depends on the materials forming the pn 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.

Considerations in use

Unlike incandescent light bulbs, which can operate with either AC or DC, LEDs require a DC supply of the correct electrical polarity. When the voltage across the pn junction is in the correct direction, a significant current flows and the device is said to be forward biased. The voltage across the LED in this case is fixed for a given LED and is proportional to the energy of the emitted photons. If the voltage is of the wrong polarity, the device is said to be reverse biased, very little current flows, and no light is emitted.

Because the voltage versus current characteristics of an LED are much like any diode, they can be destroyed by connecting them to a voltage source much higher than their turn on voltage.

The voltage drop across a forward biased LED increases as the amount of light emitted increases because of the optical power being radiated. One consequence is that LEDs of the same type can be readily operated in parallel. The turn-on voltage of an LED is a function of the color, a higher forward drop is associated with emitting higher energy (bluer) photons. The reverse voltage that most LEDs can sustain without damage is usually only a few volts. Some LED units contain two diodes, one in each direction and each a different color (typically red and green), allowing two-color operation or a range of colors to be created by altering the percentage of time the voltage is in each polarity.

LED materials

LED development began with infrared and red devices made with gallium arsenide. Advances in materials science have made possible the production of devices with ever shorter wavelengths, producing light in a variety of colors.

Conventional LEDs are made from a variety of inorganic minerals, producing the following colors:

- aluminium gallium arsenide *(AlGaAs)*: red and infrared
- gallium arsenide/phosphide *(GaAsP)*: red, orange-red, orange, and yellow
- gallium nitride *(GaN)*: green, pure green (or emerald green), and blue
- gallium phosphide *(GaP)*: red, yellow and green
- zinc selenide *(ZnSe)*: blue

- indium gallium nitride *(InGaN)*: bluish-green and blue
- silicon carbide *(SiC)*: blue
- diamond *(C)*: ultraviolet
- silicon *(Si)* - under development

(NOTE TO SELF: The above list is taken from public sources, but at least one LED given as blue does not produce blue light. (There is a good chance that almost none do, because of the higher frequency of blue.) This is a common problem in daily life due to the majority of mankind being ignorant of colour theory and conflating blue with light blue with cyan, the latter often called "sky blue". A cyan LED may be distinguished from a blue LED in that adding a yellow phosphor to the output makes green, rather than white light. And often aqua is called blue-green when in actuality the latter is cyan, and light cyan-green would be aqua. What adds to the confusion is that cyan LEDs are enclosed in blue plastic. A great amount of work is needed to dispel these intuitive myths of colour mixing before accurate descriptions of physical phenomena and their production can happen. - This needs to be sorted out)

Blue and white LEDs and Other colors

Commercially viable blue LEDs based invented by Shuji Nakamura while working in Japan at Nichia Corporation in 1993 and became widely available in the late 1990s. They can be added to existing red and green LEDs to produce white light.

Most "white" LEDs in production today use a 450nm 470nm blue GaN (gallium nitride) LED covered by a yellowish phosphor coating usually made of cerium doped yttrium aluminium garnet (YAG:Ce) crystals which have been powdered and bound in a type of viscous adhesive. The LED chip emits blue light, part of which is converted to yellow by the YAG:Ce. The single crystal form of YAG:Ce is actually considered a scintillator rather than a phosphor. Since yellow light stimulates the red and green receptors of the eye, the resulting mix of blue and yellow light gives the appearance of white.

The newest method used to produce white light LEDs uses no phosphors at all and is based on homoepitaxially grown zinc selenide (ZnSe) on a ZnSe substrate which simultaneously emits blue light from its active region and yellow light from the substrate.

Other colors

Recent color developments include pink and purple. They consist of one or two phosphor layers over a blue LED chip. The first phosphor layer of a pink LED is a yellow glowing one, and the second phosphor layer is either red or orange glowing. Purple LEDs are blue LEDs with an orange glowing phosphor over the chip. Some pink LEDs have run into issues. For example, some are blue LEDs painted with fluorescent paint or fingernail polish that can wear off, and some are white LEDs with a pink phosphor or dye that unfortunately fades after a short tme.

Ultraviolet, blue, pure green, white, pink and purple LEDs are relatively expensive compared to the more common reds, oranges, greens, yellows and infrareds and are thus less commonly used in commercial applications.

The semiconducting chip is encased in a solid plastic lens, which is much tougher than the glass envelope of a traditional light bulb or tube. The plastic may be colored, but this is only for cosmetic reasons and does not affect the color of the light emitted.

Operational parameters and efficiency

Most typical LEDs are designed to operate with no more than 30-60 milliwatts of electrical power. It is projected that by 2005, 10-watt units will be available. These devices will produce about as much light as a common 50-watt incandescent bulb, and will facilitate use of LEDs for general illumination needs.

Interesting Fact: In September 2003 a new type of blue LED was demonstrated by the company Cree, Inc. to have 35% efficiency at 20 mA. This produced a commercially packaged white light having 65 lumens per watt at 20 mA, becoming the brightest white LED commercially available at the time.

Organic light-emitting diodes (OLEDs)

If the emissive layer material of an LED is an organic compound, it is known as an Organic Light Emitting Diode (OLED). To function as a semiconductor, the organic emissive material must have conjugated pi bonds. The emissive material can be a small organic molecule in a crystalline phase, or a polymer. Polymer materials can be flexible; such LEDs are known as PLEDs or FLEDs.

Compared with regular LEDs, OLEDs are lighter and polymer LEDs can have the added benefit of being flexible. Some possible future applications of OLEDs could be:

- Light sources
- Wall decorations
- Luminous cloth

LED applications

Here is a list of known applications for LEDs, some of which are further elaborated upon in the following text:

- in general, commonly used as information indicators in various types of embedded systems (many of which are listed below)
- 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 in traffic signals, replacing ordinary bulbs behind colored glass
- car indicator lights and bicycle lighting; also for pedestrians to be seen by car traffic

- 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 backlights 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.

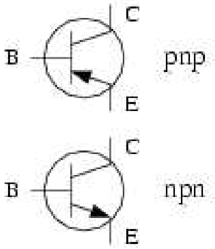
Interesting Fact: The largest LED display in the world is 36 metres high (118 feet), at Times Square, New York, U.S.A.

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.

16.3.3 Transistor

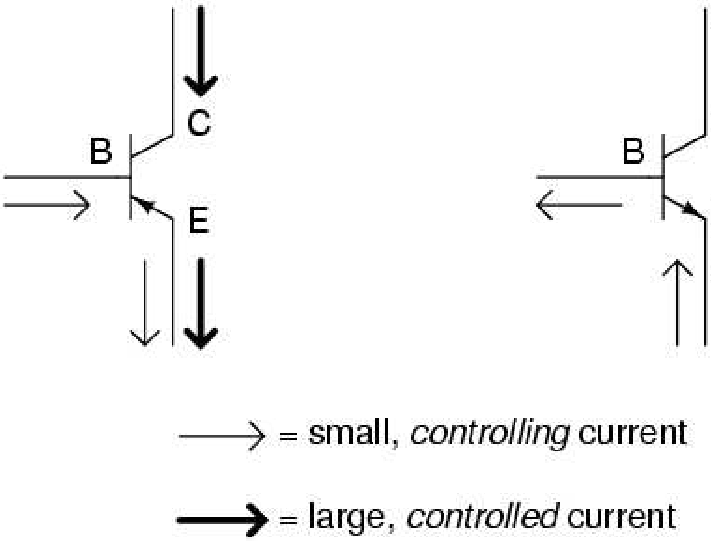
The transistor is a solid state semiconductor device used for amplification and switching, and has three terminals. The transistor itself does not amplify current though, which is a common misconception, but a small current or voltage applied to one terminal controls the current through the other two, hence the term transistor; a voltage- or current-controlled resistor. *It is the key component in all modern electronics.* In digital circuits, transistors are used as very fast electrical switches, and arrangements of transistors can function as logic gates, RAM-type memory and other devices. In analog circuits, transistors are essentially used as amplifiers.

Transistor was also the common name in the sixties for a transistor radio, a pocket-sized portable radio that utilized transistors (rather than vacuum tubes) as its active electronics. This is still one of the dictionary definitions of transistor.



The only functional difference between a PNP transistor and an NPN transistor is the proper biasing (polarity) of the junctions when operating. For any given state of operation, the current directions and voltage polarities for each type of transistor are exactly opposite each other.

Bipolar transistors work as current-controlled current regulators. In other words, they restrict the amount of current that can go through them according to a smaller, controlling current. The main current that is controlled goes from collector to emitter, or from emitter to collector, depending on the type of transistor it is (PNP or NPN, respectively). The small current that controls the main current goes from base to emitter, or from emitter to base, once again depending on the type of transistor it is (PNP or NPN, respectively). According to the confusing standards of semiconductor symbology, the arrow always points against the direction of electron flow:



Bipolar transistors are called bipolar because the main flow of electrons through them takes place in two types of semiconductor material: P and N, as the main current goes from emitter to collector (or visa-versa). In other words, two types of charge carriers – electrons and holes – comprise this main current through the transistor.

As you can see, the controlling current and the controlled current always mesh together through the emitter wire, and their electrons always flow against the direction of the transistor's arrow. This is the first and foremost rule in the use of transistors: all currents must be going in the proper directions for the device to work as a current regulator. The small, controlling current is usually referred to simply as the base current because it is the only current that goes through the base wire of the transistor. Conversely, the large, controlled current is referred to as the collector current because it is the only current that goes through the collector wire. The emitter current is the sum of the base and collector currents, in compliance with Kirchhoff's Current Law.

If there is no current through the base of the transistor, it shuts off like an open switch and prevents current through the collector. If there is a base current, then the transistor turns on like a closed switch and allows a proportional amount of current through the collector. Collector current is primarily limited by the base current, regardless of the amount of voltage available to push it. The next section will explore in more detail the use of bipolar transistors as switching elements.

Importance

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 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.

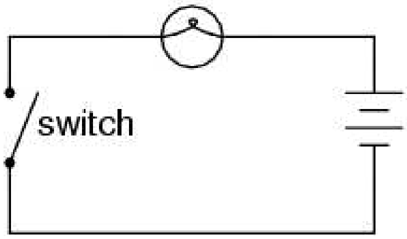
Hand-in-hand with low cost has been the increasing move to "digitizing" all information. With transistorized computers offering the ability to quickly find (and sort) digital information, more and more effort was put into making all information digital. Today almost all media in modern society is delivered in digital form, converted and presented by computers. Common "analog" forms of information such as television or newspapers spend the vast majority of their time as digital information, being converted to analog only for a small portion of the time.

Interesting Fact: 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.

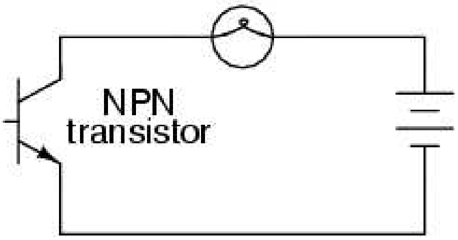
16.3.4 The transistor as a switch

Because a transistor's collector current is proportionally limited by its base current, it can be used as a sort of current-controlled switch. A relatively small flow of electrons sent through the base of the transistor has the ability to exert control over a much larger flow of electrons through the collector.

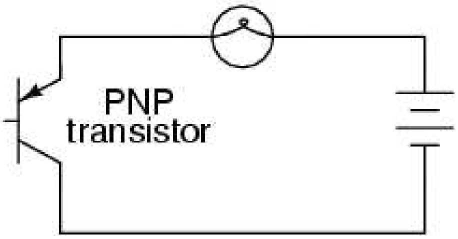
Suppose we had a lamp that we wanted to turn on and off by means of a switch. Such a circuit would be extremely simple:



For the sake of illustration, let's insert a transistor in place of the switch to show how it can control the flow of electrons through the lamp. Remember that the controlled current through a transistor must go between collector and emitter. Since it's the current through the lamp that we want to control, we must position the collector and emitter of our transistor where the two contacts of the switch are now. We must also make sure that the lamp's current will move against the direction of the emitter arrow symbol to ensure that the transistor's junction bias will be correct:

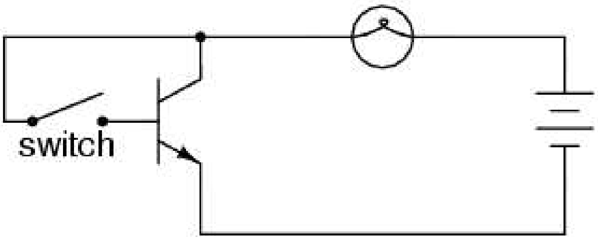


In this example I happened to choose an NPN transistor. A PNP transistor could also have been chosen for the job, and its application would look like this:

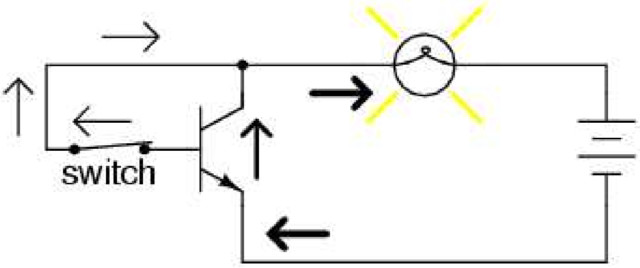


The choice between NPN and PNP is really arbitrary. All that matters is that the proper current directions are maintained for the sake of correct junction biasing (electron flow going against the transistor symbol's arrow).

Going back to the NPN transistor in our example circuit, we are faced with the need to add something more so that we can have base current. Without a connection to the base wire of the transistor, base current will be zero, and the transistor cannot turn on, resulting in a lamp that is always off. Remember that for an NPN transistor, base current must consist of electrons flowing from emitter to base (against the emitter arrow symbol, just like the lamp current). Perhaps the simplest thing to do would be to connect a switch between the base and collector wires of the transistor like this:

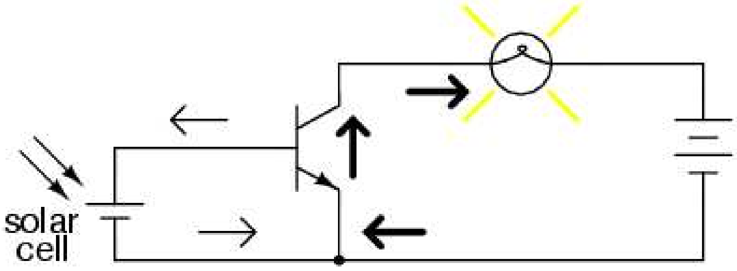


If the switch is open, the base wire of the transistor will be left "floating" (not connected to anything) and there will be no current through it. In this state, the transistor is said to be cutoff. If the switch is closed, however, electrons will be able to flow from the emitter through to the base of the transistor, through the switch and up to the left side of the lamp, back to the positive side of the battery. This base current will enable a much larger flow of electrons from the emitter through to the collector, thus lighting up the lamp. In this state of maximum circuit current, the transistor is said to be saturated.

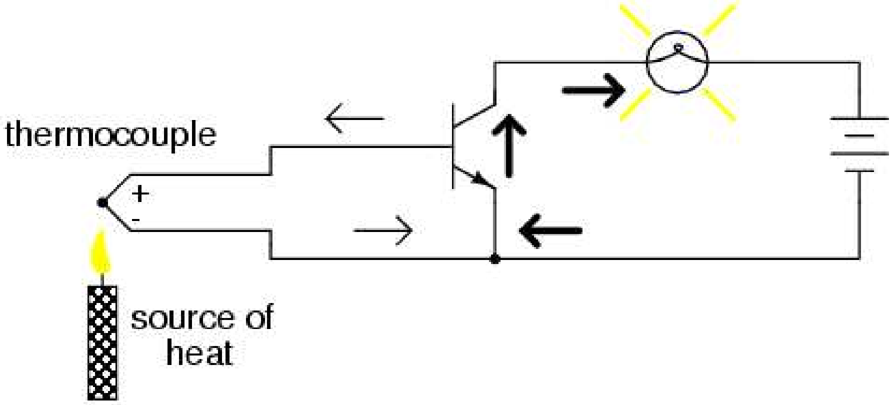


Of course, it may seem pointless to use a transistor in this capacity to control the lamp. After all, we're still using a switch in the circuit, aren't we? If we're still using a switch to control the lamp – if only indirectly – then what's the point of having a transistor to control the current? Why not just go back to our original circuit and use the switch directly to control the lamp current?

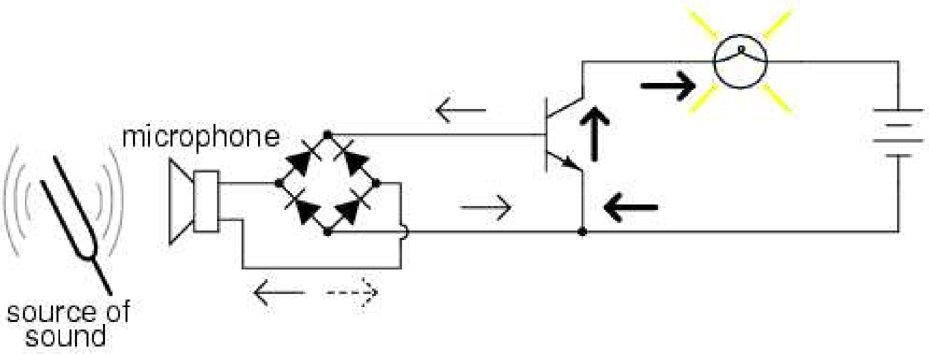
There are a couple of points to be made here, actually. First is the fact that when used in this manner, the switch contacts need only handle what little base current is necessary to turn the transistor on, while the transistor itself handles the majority of the lamp's current. This may be an important advantage if the switch has a low current rating: a small switch may be used to control a relatively high-current load. Perhaps more importantly, though, is the fact that the current-controlling behavior of the transistor enables us to use something completely different to turn the lamp on or off. Consider this example, where a solar cell is used to control the transistor, which in turn controls the lamp:



Or, we could use a thermocouple to provide the necessary base current to turn the transistor on:



Even a microphone of sufficient voltage and current output could be used to turn the transistor on, provided its output is rectified from AC to DC so that the emitter-base PN junction within the transistor will always be forward-biased:



The point should be quite apparent by now: any sufficient source of DC current may be used to turn the transistor on, and that source of current need only be a fraction of the amount of current needed to energize the lamp. Here we see the transistor functioning not only as a switch, but as a true amplifier: using a relatively low-power signal to control a relatively large amount of power. Please note that the actual power for lighting up the lamp comes from the battery to the right of the schematic. It is not as though the small signal current from the solar cell, thermocouple, or microphone is being magically transformed into a greater amount of power. Rather, those small power sources are simply controlling the battery's power to light up the lamp.

Field-Effect Transistor (FET)

(NOTE TO SELF: Schematic can be found under GFDL on wikipedia)

The schematic symbols for p- and n-channel MOSFETs. The symbols to the right include an extra terminal for the transistor body (allowing for a seldom-used channel bias) whereas in those to the left the body is implicitly connected to the source.

The most common variety of field-effect transistors, the enhancement-mode MOSFET (metal-oxide semiconductor field-effect transistor) consists of a unipolar conduction channel and a metal gate separated from the main conduction channel by a thin layer of (SiO2) glass. This is why an alternative name for the FET is 'unipolar transistor.' When a potential difference (of the proper polarity) is impressed across gate and source, charge carriers are introduced to the channel, making it conductive. The amount of this current can be modulated, or (nearly) completely turned off, by varying the gate potential.

Because the gate is insulated, no DC current flows to or from the gate electrode. This lack of a gate current and the ability of the MOSFET to act like a switch, allows particularly efficient digital circuits to be created, with very low power consumption at low frequencies. The power consumption increases markedly with frequency, because the capacitive loading of the FET control terminal takes more energy to slew at higher frequencies, in direct proportion to the frequency. Hence, MOSFETs have become the dominant technology used in computing hardware such as microprocessors and memory devices such as RAM. Bipolar transistors are more rugged and hence more useful for low-impedance loads and inductively reactive (e.g. motor) loads.

Power MOSFETs become less conductive with increasing temperature and can therefore be applied in shunt, to increase current capacity, unlike the bipolar transistor, which has a negative temperature coefficient of resistance, and is therefore prone to thermal runaway. The downside of this is that, while the power FET can protect itself from overheating by diminishing the current through it, high temperatures need to be avoided by using a larger heat sink than for an equivalent bipolar device. Macroscopic FET power transistors are actually composed of many little transistors. They are stacked (on-chip) to increase breakdown potential and paralleled to reduce Ron, i.e. allowing for more current, bussing the gates to provide a single control (gate) terminal.

The depletion mode FET is a little different. It uses a back-biased diode for the control terminal, which presents a capacitive load to the driving circuit in normal operation. With the gate tied to the source, a DFET is fully on. Changing the potential of a DFET (pulling an N-channel gate downward, for example) will turn it off, i.e. 'deplete' the channel (drain-source) of charge carriers. MOSFETs, formerly called IGFETs (for Insulated Gate Field-Effect Transistor) can be depletion-mode, enhancement-mode, or mixed-mode, but are almost always enhancement mode in modern commercial practice. This means that, with the source and gate tied together (thus equipotential) the channel will be off (high impedance or non-conducting). The n-channel device (reverse for P-channel), like in the DFET, is turned on by raising the potential of the gate. Typically, the gate on a MOSFET will withstand +-20V, relative to the source terminal. If one were to raise the gate potential of an n-channel device without limiting the current to a few milliamps, one would destroy the gate diode, like any other small diode. Why do we typically think of n-channel devices as the default? In silicon devices, the ones that use majority carriers that are electrons, rather than holes, are slightly faster and can carry more current than their P-type counterparts. The same is true in GaAs devices.

The FET is simpler in concept than the bipolar transistor and can be constructed from a wide range of materials.

The most common use of MOSFET transistors today is the CMOS (complementary metallic oxide semiconductor) integrated circuit which is the basis for most digital electronic devices. These use a totem-pole arrangement where one transistor (either the pull-up or the pull-down) is on while the other is off. Hence, there is no DC drain, except during the transition from one state to the other, which is very short. As mentioned, the gates are capacitive, and the charging and discharging of the gates each time a transistor switches states is the primary cause of power drain.

The C in CMOS stands for 'complementary.' The pull-up is a P-channel device (using holes for the mobile carrier of charge) and the pull-down is N-channel (electron carriers). This allows busing of the control terminals, but limits the speed of the circuit to that of the slower P device (in silicon devices). The bipolar solutions to push-pull include 'cascode' using a current source for the load. Circuits that utilize both unipolar and bipolar transistors are called Bi-Fet. A recent development is called 'vertical P.' Formerly, BiFet chip users had to settle for relatively poor (horizontal) P-type FET devices. This is no longer the case and allows for quieter and faster analog circuits.

A clever variant of the FET is the dual-gate device. This allows for two opportunities to turn the device off, as opposed to the dual-base (bipolar) transistor which presents two opportunities to turn the device on.

FETs can switch signals of either polarity, if their amplitude is significantly less than the gate swing, as the devices (especially the parasitic diode-free DFET) are basically symmetrical. This means that FETs are the most suitable type for analog multiplexing. With this concept, one can construct a solid-state mixing board, for example.

The power MOSFET has a 'parasitic diode' (back-biased) normally shunting the conduction channel that has half the current capacity of the conduction channel. Sometimes this is useful in driving dual-coil magnetic circuits (for spike protection), but in other cases it causes problems.

The high impedance of the FET gate makes it rather vulnerable to electrostatic damage, though this is not usually a problem after the device has been installed.

A more recent device for power control is the insulated-gate bipolar transistor, or IGBT. This has a control structure akin to a MOSFET coupled with a bipolar-like main conduction channel. These have become quite popular.

16.4 principles of digital electronics logical gates, counting circuits

16.4.1 Electronic logic gates

The simplest form of electronic logic is diode logic (DL). This allows AND and OR gates to be built, but not inverters, and so is an incomplete form of logic. To built a complete logic system, valves or transistors can be used. The simplest family of logic gates using bipolar transistors is called resistor-transistor logic, or RTL. Unlike diode logic gates, RTL gates can be cascaded indefinitely to produce more complex logic functions. These gates were used in early integrated circuits. For higher speed, the resistors used in RTL were replaced by diodes, leading to diode-transistor logic, or DTL. It was then discovered that one transistor could do the job of two diodes in the space of one diode, so transistor-transistor logic, or TTL, was created. In some types of chip, to reduce size and power consumption still further, the bipolar transistors were replaced with complementary field-effect transistors (MOSFETs), resulting in complementary metal-oxide-semiconductor (CMOS) logic.

For small-scale logic, designers now use prefabricated logic gates from families of devices such as the TTL 7400 series invented by Texas Instruments and the CMOS 4000 series invented by RCA, and their more recent descendants. These devices usually contain transistors with multiple emitters, used to implement the AND function, which are not available as separate components. Increasingly, these fixed-function logic gates are being replaced by programmable logic devices, which allow designers to pack a huge number of mixed logic gates into a single integrated circuit.

Electronic logic gates differ significantly from their relay-and-switch equivalents. They are much faster, consume much less power, and are much smaller (all by a factor of a million or more in most cases). Also, there is a fundamental structural difference. The switch circuit creates a continuous metallic path for current to flow (in either direction) between its input and its output. The semiconductor logic gate, on the other hand, acts as a high-gain voltage amplifier, which sinks a tiny current at its input and produces a low-impedance voltage at its output. It is not possible for current to flow between the output and the input of a semiconductor logic gate.

Another important advantage of standardised semiconductor logic gates, such as the 7400 and 4000 families, is that they are cascadable. This means that the output of one gate can be wired to the inputs of one or several other gates, and so on ad infinitum, enabling the construction of circuits of arbitrary complexity without requiring the designer to understand the internal workings of the gates.

In practice, the output of one gate can only drive a finite number of inputs to other gates, a number called the 'fanout limit', but this limit is rarely reached in the newer CMOS logic circuits, as compared to TTL circuits. Also, there is always a delay, called the 'propagation delay', from a change an input of a gate to the corresponding change in its output. When gates are cascaded, the total propagation delay is approximately the sum of the individual delays, an effect which can become a problem in high-speed circuits.

The US symbol for an AND gate is: AND symbol and the IEC symbol is AND symbol.

The US circuit symbol for an OR gate is: OR symbol and the IEC symbol is: OR symbol.

The US circuit symbol for a NOT gate is: NOT symbol and the IEC symbol is: NOT symbol.

In electronics a NOT gate is more commonly called an inverter. The circle on the symbol is called a bubble, and is generally used in circuit diagrams to indicate an inverted input or output.

The US circuit symbol for a NAND gate is: NAND symbol and the IEC symbol is: NAND symbol.

The US circuit symbol for a NOR gate is: NOR symbol and the IEC symbol is: NOR symbol.

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.

Two more gates are the exclusive-OR or XOR function and its inverse, exclusive-NOR or XNOR. Exclusive-OR is true only when exactly one of its inputs is true. In practice, these gates are built from combinations of simpler logic gates.

The US circuit symbol for an XOR gate is: XOR symbol and the IEC symbol is: XOR symbol.

16.5 Counting circuits

An arithmetic and logical unit (ALU) adder provides the basic functionality of arithmetic operations within a computer, and is a significant component of the arithmetic and logical unit. Adders are composed of half adders and full adders, which add two-bit binary pairs, and ripple carry adders and carry look ahead adders which do addition operations to a series of binary numbers.

(NOTE TO SELF: Pictures on wikipedia under GFDL)

16.5.1 Half Adder

A half adder is a logical circuit that performs an addition operation on two binary digits. The half adder produces a sum and a carry value which are both binary digits.

Sum(s) = A xor B Cot(c) = A and B

Half adder circuit diagram Half adder circuit diagram

Following is the logic table for a half adder:

A	B	Sum	Cot
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

16.5.2 Full adder

A full adder is a logical circuit that performs an addition operation on three binary digits. The full adder produces a sum and carry value, which are both binary digits.

Sum = (A xor B) xor Cin Cot = (A nand B) nand (Cin nand (A xor B))

Full adder circuit diagram Full adder circuit diagram

A	B	Cin	Sum	Cot
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

Units				
Quantity	Symbol	Unit	S.I. Units	Direction
			or	

Table 16.1: Units used in **Electronics**

Chapter 17

The Atom

Atoms are the building blocks of matter. They are the basis of all the structures and organisms in the universe. The planets, the sun, grass and trees, the air we breathe, and people are all made up of atoms.

17.1 Models of the Atom

17.2 Structure of the Atom

Atoms are very small and cannot be seen with the naked eye. They consist of two main parts: the positively charged **nucleus** at the centre and the negatively charged elementary particles called **electrons** which surround the nucleus in their **orbitals**. (*Elementary* particle means that the electron cannot be broken down to anything smaller and can be thought of as a point particle.) The nucleus of an atom is made up of a collection of positively charged **protons** and neutral particles called **neutrons**.

interesting fact: the neutrons and protons are *not* elementary particles. They are actually made up of even smaller particles called quarks. Both protons and neutrons are made of three quarks each. There are all sorts of other particles composed of quarks which nuclear physicists study using huge detectors - you can find out more about this by reading the essay in Chapter ??.

(NOTE TO SELF: Insert diagram of atomic structure - see lab posters)

Atoms are electrically neutral which means that they have the same number of negative electrons as positive protons. The number of protons in an atom is called the **atomic number** which is sometimes also called **Z**. (NOTE TO SELF: check A and Z) The atomic number is what distinguishes the different chemical elements in the Periodic table from each other. In fact, the elements are listed on the Periodic table in order of their atomic numbers. For example, the first element, hydrogen (H), has one proton whereas the sixth element, carbon (C) has 6 protons. Atoms with the same number of protons (atomic number) share physical properties and show similar chemical behaviour. The number of neutrons *plus* protons in the nucleus is called the **atomic mass** of the atom.

17.3 Isotopes

Two atoms are considered to be the same element if they have the same number of protons (atomic number). However, they do not have to have the same number of neutrons or overall atomic mass. Atoms which have the same number of protons but different numbers of neutrons are called **isotopes**. For example, the hydrogen atom has one proton and no neutrons. Therefore its atomic number is Z=1 and atomic mass is A=1. If a neutron is added to the hydrogen nucleus, then a new atom is formed with atomic mass A=2 but atomic number is still Z=1. This atom is called deuterium and is an isotope of hydrogen.

17.4 Energy quantization and electron configuration

17.5 Periodicity of ionization energy to support atom arrangement in Periodic Table

17.6 Successive ionisation energies to provide evidence for arrangement of electrons into core and valence

[Brink and Jones sections: de Broglie - matter shows particle and wave characteristics, proved by Davisson and Germer. Shroedinger and Heisenberg developed this model into quantum mechanics]

The nucleus (atomic nucleus) is the center of an atom. It is composed of one or more protons and usually some neutrons as well. The number of protons in an atom's nucleus is called the atomic number, and determines which element the atom is (for example hydrogen, carbon, oxygen, etc.).

Though the positively charged protons exert a repulsive electromagnetic force on each other, the distances between nuclear particles are small enough that the strong interaction (which is stronger than the electromagnetic force but decreases more rapidly with distance) predominates. (The gravitational attraction is negligible, being a factor 1036 weaker than this electromagnetic repulsion.)

The discovery of the electron was the first indication that the atom had internal structure. This structure was initially imagined according to the ”raisin cookie” or ”plum pudding” model, in which the small, negatively charged electrons were embedded in a large sphere containing all the positive charge. Ernest Rutherford and Marsden, however, discovered in 1911 that alpha particles from a radium source were sometimes scattered backwards from a gold foil, which led to the acceptance of a planetary model, in which the electrons orbited a tiny nucleus in the same way that the planets orbit the sun.

Interesting Fact: The word atom is derived from the Greek atomos, indivisible, from a-, not, and tomos, a cut.

An atom is the smallest portion into which a chemical element can be divided while still retaining its properties. Atoms are the basic constituents of molecules and ordinary matter. Atoms are composed of subatomic particles.

Atoms are composed mostly of empty space, but also of smaller subatomic particles. At the center of the atom is a tiny positive nucleus composed of nucleons (protons and neutrons). The rest of the atom contains only the fairly flexible electron shells. Usually atoms are electrically neutral with as many electrons as protons.

Atoms are generally classified by their atomic number, which corresponds to the number of protons in the atom. For example, carbon atoms are those atoms containing 6 protons. All atoms with the same atomic number share a wide variety of physical properties and exhibit the same chemical behavior. The various kinds of atoms are listed in the Periodic table. Atoms having the same atomic number, but different atomic masses (due to their different numbers of neutrons), are called isotopes.

The simplest atom is the hydrogen atom, having atomic number 1 and consisting of one proton and one electron. It has been the subject of much interest in science, particularly in the early development of quantum theory.

The chemical behavior of atoms is largely due to interactions between the electrons. In particular the electrons in the outermost shell, called the valence electrons, have the greatest influence on chemical behavior. Core electrons (those not in the outer shell) play a role, but it is usually in terms of a secondary effect due to screening of the positive charge in the atomic nucleus.

There is a strong tendency for atoms to completely fill (or empty) the outer electron shell, which in hydrogen and helium has space for two electrons, and in all other atoms has space for eight. This is achieved either by sharing electrons with neighboring atoms or by completely removing electrons from other atoms. When electrons are shared a covalent bond is formed between the two atoms. Covalent bonds are the strongest type of atomic bond.

When one or more electrons are completely removed from one atom by another, ions are formed. Ions are atoms that possess a net charge due to an imbalance in the number of protons and electrons. The ion that stole the electron(s) is called an anion and is negatively charged. The atom that lost the electron(s) is called a cation and is positively charged. Cations and anions are attracted to each other due to coulombic forces between the positive and negative charges. This attraction is called ionic bonding and is weaker than covalent bonding.

As mentioned above covalent bonding implies a state in which electrons are shared equally between atoms, while ionic bonding implies that the electrons are completely confined to the anion. Except for a limited number of extreme cases, neither of these pictures is completely accurate. In most cases of covalent bonding, the electron is unequally shared, spending more time around the more electronegative atom, resulting in the covalent bond having some ionic character. Similarly, in ionic bonding the electrons often spend a small fraction of time around the more electropositive atom, resulting in some covalent character for the ionic bond. [edit]

Models of the atom

* Democritus' shaped-atom model (for want of a better name) * The plum pudding model * Cubical atom * The Bohr model * The quantum mechanical model

The Plum pudding model of the atom was made after the discovery of the electron but before the discovery of the proton or neutron. In it, the atom is envisioned as electrons surrounded by a soup of positive charge, like plums surrounded by pudding. This model was disproved by an experiment by Ernest Rutherford when he discovered the nucleus of the atom.

The Bohr Model is a physical model that depicts the atom as a small positively charged nucleus with electrons in orbit at different levels, similar in structure to the solar system. Because of its simplicity, the Bohr model is still commonly used and taught today.

In the early part of the 20th century, experiments by Ernest Rutherford and others had established that atoms consisted of a small dense positively charged nucleus surrounded by orbiting negatively charged electrons. However classical physics at that time was unable to explain why the orbiting electrons did not spiral into the nucleus.

The simplest possible atom is hydrogen, which consists of a nucleus and one orbiting electron. Since the nucleus is positive and the electron are oppositely charged they will attract one another by coulomb force, in much the same way that the sun attracts the earth by gravitational force.

However, if the electron orbits the nucleus in a classical orbit, it ought to emit electromagnetic radiation (light) according to well established theories of electromagnetism.

If the orbiting electron emits light, it must lose energy and spiral into the nucleus, so why do atoms even exist? What's more, the spectra of atoms show that the orbiting electrons can emit light but only at certain frequencies. This made no sense at all to the scientists of the time.

These difficulties were resolved in 1913 by Niels Bohr who proposed that:

* (1) The orbiting electrons existed in orbits that had discrete quantized energies. That is, not every orbit is possible but only certain specific ones. The exact energies of the allowed orbits depends on the atom in question. * (2) The laws of classical mechanics do not apply when electrons make the jump from one allowed orbit to another. * (3) When an electron makes a jump from one orbit to another the energy difference is carried off (or supplied) by a single quantum of light (called a photon) which has a frequency that directly depends on the energy difference between the two orbitals.

f = E / h

where f is the frequency of the photon, E the energy difference, and h is a constant of proportionality known as Planck's constant. Defining we can write

where ? is the angular frequency of the photon.

* (4) The allowed orbits depend on quantized (discrete) values of orbital angular momentum, L according to the equation

Where n = 1,2,3, and is called the angular momentum quantum number.

These assumptions explained many of the observations seen at the time, such as why spectra consist of discrete lines. Assumption 4) states that the lowest value of n is 1. This corresponds to a smallest possible radius (for the mathematics see Ohanian-principles of physics or any of the large, usually American, college introductory physics textbooks) of 0.0529 nm. This is known as the Bohr radius, and explains why atoms are stable. Once an electron is in the lowest orbit, it can go no further. It cannot emit any more light because it would need to go into a lower orbit, but it can't do that if it is already in the lowest allowed orbit.

The Bohr model is sometimes known as the semiclassical model because although it does include some ideas of quantum mechanics it is not a full quantum mechanical description of the atom. Assumption 2) states that the laws of classical mechanics don't apply during a quantum jump but doesn't state what laws should replace classical mechanics. Assumption 4) states that angular momentum is quantised but does not explain why.

In order to fully describe an atom we need to use the full theory of quantum mechanics, which was worked out by a number of people in the years following the Bohr model. This theory treats the electrons as waves, which create 3D standing wave patterns in the atom. (This is why quantum mechanics is sometimes called wave mechanics.) This theory considers that idea of electrons as being little billiard ball like particles that travel round in orbits as absurdly wrong; instead electrons form probability clouds. You might find the electron here with a certain probability; you might find it over there with a different probability. However it is interesting to note that if you work out the average radius of an electron in the lowest possible energy state it turns out to be exactly equal to the Bohr radius (although it takes many more pages of mathematics to work it out).

The full quantum mechanics theory is a beautiful theory that has been experimentally tested and found to be incredibly accurate, however it is mathematically much more advanced, and often using the much simpler Bohr model will get you the results with much less hassle. The thing to remember is that it is only a model, an aid to understanding. Atoms are not really little solar systems.

* See also: Hydrogen atom, quantum mechanics, Schrdinger equation, Niels Bohr. * An interactive demonstration (http://webphysics.davidson.edu/faculty/dmb/hydrogen/) of the prob-

ability clouds of electron in Hydrogen atorm according to the full QM solution.

17.7 Bohr orbits

Brink and Jones sections: Standing waves (quantisation). Atom seen as positive nucleus with vibrating electron waves surrounding it. Shrodinger's equation calucaltes the energy of these waves and their shape and position– most probable region of movement of electrons called orbitals (talk about n=1,2 energy levels and spdf orbitals).

17.8 Heisenberg uncertainty Principle

Quantum mechanics is a physical theory that describes the behavior of physical systems at short distances. Quantum mechanics provides a mathematical framework derived from a small set of basic principles capable of producing experimental predictions for three types of phenomena that classical mechanics and classical electrodynamics cannot account for: quantization, wave-particle duality, and quantum entanglement. The related terms quantum physics and quantum theory are sometimes used as synonyms of quantum mechanics, but also to denote a superset of theories, including pre-quantum mechanics old quantum theory, or, when the term quantum mechanics is used in a more restricted sense, to include theories like quantum field theory.

Quantum mechanics is the underlying theory of many fields of physics and chemistry, including condensed matter physics, quantum chemistry, and particle physics.

17.9 Pauli exclusion principle

The Pauli exclusion principle is a quantum mechanical principle which states that no two identical fermions may occupy the same quantum state. Formulated by Wolfgang Pauli in 1925, it is also referred to as the "exclusion principle" or "Pauli principle."

The Pauli principle only applies to fermions, particles which form antisymmetric quantum states and have half-integer spin. Fermions include protons, neutrons, and electrons, the three types of elementary particles which constitute ordinary matter. The Pauli exclusion principle governs many of the distinctive characteristics of matter. Particles like the photon and graviton do not obey the Pauli exclusion principle, because they are bosons (i.e. they form symmetric quantum states and have integer spin) rather than fermions.

The Pauli exclusion principle plays a role in a huge number of physical phenomena. One of the most important, and the one for which it was originally formulated, is the electron shell structure of atoms. An electrically neutral atom contains bound electrons equal in number to the protons in the nucleus. Since electrons are fermions, the Pauli exclusion principle forbids them from occupying the same quantum state.

For example, consider a neutral helium atom, which has two bound electrons. Both of these electrons can occupy the lowest-energy (1s) states by acquiring opposite spin. This does not violate the Pauli principle because spin is part of the quantum state of the electron, so the two electrons are occupying different quantum states. However, the spin can take only two different values (or eigenvalues.) In a lithium atom, which contains three bound electrons, the third electron cannot fit into a 1s state, and has to occupy one of the higher-energy 2s states instead. Similarly, successive elements produce successively higher-energy shells. The chemical properties of an element largely depends on the number of electrons in the outermost shell, which gives rise to the periodic table of the elements.

The Pauli principle is also responsible for the large-scale stability of matter. Molecules cannot be pushed arbitrarily close together, because the bound electrons in each molecule are forbidden from entering the same state as the electrons in the other molecules - this is the reason for the repulsive r-12 term in the Lennard-Jones potential. The Pauli principle is the reason you do not fall through the floor.

Astronomy provides the most spectacular demonstrations of this effect, in the form of white dwarf stars and neutron stars. In both types of objects, the usual atomic structures are disrupted by large gravitational forces, leaving the constituents supported only by a "degeneracy pressure" produced by the Pauli exclusion principle. This exotic form of matter is known as degenerate matter. In white dwarfs, the atoms are held apart by the degeneracy pressure of the electrons. In neutron stars, which exhibit even larger gravitational forces, the electrons have merged with the protons to form neutrons, which produce a larger degeneracy pressure.

Another physical phenomenon for which the Pauli principle is responsible is ferromagnetism, in which the exclusion effect implies an exchange energy that induces neigboring electron spins to align (whereas classically they would anti-align).

17.10 Ionization Energy

(first, second etc.)

17.11 Electron configuration

i.e. filling the orbitals starting from 1s..... Aufbau principle unpaired and paired electrons Hund's rule: 1 e- in each orbital before pairing in p orbitals shorthand: $1s^2 2s^2 2p^1$ etc

17.12 Valency

Capacity for bonding

Covalent bonding is a form of chemical bonding characterized by the sharing of one or more pairs of electrons, by two atoms, in order to produce a mutual attraction; atoms tend to share electrons, so as to fill their outer electron shells. Such bonds are always stronger than the intermolecular hydrogen bond and similar in strength or stronger than the ionic bond. Commonly covalent bond implies the sharing of just a single pair of electrons. The sharing of two pairs is called a double bond and three pairs is called a triple bond. Aromatic rings of atoms and other resonant structures are held together by covalent bonds that are intermediate between single and double. The triple bond is relatively rare in nature, and two atoms are not observed to bond more than triply.

Covalent bonding most frequently occurs between atoms with similar electronegativities, where neither atom can provide sufficient energy to completely remove an electron from the other atom. Covalent bonds are more common between non-metals, whereas ionic bonding is more common between two metal atoms or a metal and a non-metal atom.

Covalent bonding tends to be stronger than other types of bonding, such as ionic bonding. In addition unlike ionic bonding, where ions are held together by a non-directional coulombic attraction, covalent bonds are highly directional. As a result, covalently bonded molecules tend to form in a relatively small number of characteristic shapes, exhibiting specific bonding angles.

17.13

Chapter 18

Modern Physics

18.1 Introduction to the idea of a quantum

Imagine that a beam of light is actually made up of little "packets" or "bundles" of energy, called quanta.

It's like looking at a crowd of people from above. At first, it seems as though they are one huge patch, without any spaces between them. You would never suspect that they were people. But as you move closer, you slowly begin to see that they are individuals, and when you get even closer, you may even recognize a few. Light seems like a continuous wave at first, but when we zoom in at the subatomic level, we notice that a beam of light actually consists of little "packets" of energy, or quanta.

This idea introduces the concept of the quantum (particle) nature of light, which is demonstrated by the photoelectric effect. When a metal surface is illuminated with light, electrons can be emitted from the surface. This is known as the photoelectric effect.

18.2 The wave-particle duality

The wave nature of light is demonstrated by diffraction, interference, and polarization of light; and the particle nature of light is demonstrated by the photoelectric effect. So light has both wave-like and particle-like properties, but only shows one or the other, depending on the kind of experiment we perform. A wave-type experiment shows the wave nature, and a particle-type experiment shows particle nature.

When you're watching a cricketer on the field, you see only that side of his personality. So to you, he is just a good cricketer. You do not see his golfing side, for example. Only when he is playing golf, will that side be revealed to you. The same applies to light.

Now, we consider light to behave not as a wave, but as particles. But what do we call a 'particle' of light?

Photon : A photon is a quantum (energy packet) of light.

Imagine a sheet of metal. On the surface, there are electrons that are waiting to be set free. If a photon comes along and strikes the surface of the metal, then it will give its entire energy packet to one electron. This means that the electron now has some energy, and it may escape (leave the surface) if this energy Ek is greater than the minimum energy required to free an electron Emin.

Now, suppose the electron needs 5eV of kinetic energy to escape. And suppose this little photon has just 2eV of energy in its energy packet. Then the electron will not leave the surface of the metal. But suppose the photon has 8eV of energy. This means that the electron will emerge with 3eV of kinetic energy.

Note that this does not mean the photon can give 5eV of energy to one electron and 3eV to another. A photon will give all of its energy to just one electron.

The minimum amount of energy needed for an electron to escape (electrons do not normally leave a metal whenever they please), is called the work function of the metal. In our example, the work function is 5eV. The work function has a different value for each metal: 4.70eV for copper and 2.28eV for sodium. It is worth mentioning that the best conductors are those with the smallest work functions.

The frequency of the radiation is very important, because if it is below a certain threshold value, no electrons will be emitted. Even if the intensity of the light is increased, and the light is allowed to fall on the surface for a long period of time, if the frequency of the radiation is below the threshold frequency, electrons will not be emitted.

We therefore reason that E = h? Where E is the energy of the photon, $h = 6.57X10^-34Js$ is Plancks constant, and ? is the frequency of the radiation.

This means that the kinetic energy acquired by the electron is equal to the energy of the photon minus the work function, ?, i.e. Ek = h? - ? The electrons emerge with a range of velocities from zero up to a maximum vmax. The maximum kinetic energy, (1/2)mvmax2, depends (linearly) on the frequency of the radiation, and is independent of its intensity.

For incident radiation of a given frequency, the number of electrons emitted per unit time is proportional to the intensity of the radiation.

Electron emission takes place from the instant the light shines on the surface, i.e. there is no detectable time delay.

What are the uses of the photoelectric effect?

For this work, Einstein received the Nobel prize in 1905.

18.3 Practical Applications of Waves: Electromagnetic Waves

In physics, wave-particle duality holds that light and matter simultaneously exhibit properties of waves and of particles. This concept is a consequence of quantum mechanics.

In 1905, Einstein reconciled Huygens' view with that of Newton; he explained the photoelectric effect (an effect in which light did not seem to act as a wave) by postulating the existence of photons, quanta of energy with particulate qualities. Einstein postulated that light's frequency, ?, is related to the energy, E, of its photons:

$$E = hf \tag{18.1}$$

where h is Planck's constant ($6.626 \times 10^{-34} Js$).

In 1924, De Broglie claimed that all matter has a wave-like nature; he related wavelength, ?, and momentum, p:

$$\lambda = \frac{h}{p} \tag{18.2}$$

This is a generalization of Einstein's equation above since the momentum of a photon is given by,

$$p = \frac{E}{c}, \tag{18.3}$$

where c is the speed of light in vacuum, and ? = c / ?.

De Broglie's formula was confirmed three years later by guiding a beam of electrons (which have rest mass) through a crystalline grid and observing the predicted interference patterns. Similar experiments have since been conducted with neutrons and protons. Authors of similar recent experiments with atoms and molecules claim that these larger particles also act like waves. This is still a contoversial subject because these experimenters have assumed arguments of wave-particle duality and have assumed the validity of deBroglie's equation in their argument.

The Planck constant h is extremely small and that explains why we don't perceive a wave-like quality of everyday objects: their wavelengths are exceedingly small. The fact that matter can have very short wavelengths is exploited in electron microscopy.

In quantum mechanics, the wave-particle duality is explained as follows: every system and particle is described by state functions which encode the probability distributions of all measurable variables. The position of the particle is one such variable. Before an observation is made the position of the particle is described in terms of probability waves which can interfere with each other.

Chapter 19

Inside atomic nucleus

Amazingly enough, human mind that is kind of contained inside a couple of liters of human's brain, is able to deal with extremely large as well as extremely small objects such as the whole universe and its smallest building blocks. So, what are these building blocks? As we already know, the universe consists of galaxies, which consist of stars with planets moving around. The planets are made of molecules, which are bound groups (chemical compounds) of atoms.

There are more than 10^{20} stars in the universe. Currently, scientists know over 12 million chemical compounds i.e. 12 million different molecules. All this variety of molecules is made of only a hundred of different atoms. For those who believe in beauty and harmony of nature, this number is still too large. They would expect to have just few different things from which all other substances are made. In this chapter, we are going to find out what these elementary things are.

19.1 What the atom is made of

The Greek word $\alpha\tau o\mu o\nu$ (atom) means indivisible. The discovery of the fact that an atom is actually a complex system and can be broken in pieces was the most important step and pivoting point in the development of modern physics.

It was discovered (by Rutherford in 1911) that an atom consists of a positively charged nucleus and negative electrons moving around it. At first, people tried to visualize an atom as a microscopic analog of our solar system where planets move around the sun. This naive planetary model assumes that in the world of very small objects the Newton laws of classical mechanics are valid. This, however, is not the case.

The microscopic world is governed by quantum mechanics which does not have such notion as trajectory. Instead, it describes the dynamics of particles in terms of quantum states that are characterized by probability distributions of various observable quantities.

For example, an electron in the atom is not moving along a certain trajectory but rather along all imaginable trajectories with different probabilities. If we were trying to catch this electron, after many such attempts we would discover that the electron can be found anywere around the nucleus, even very close to and very far from it. However, the probabilities of finding the electron

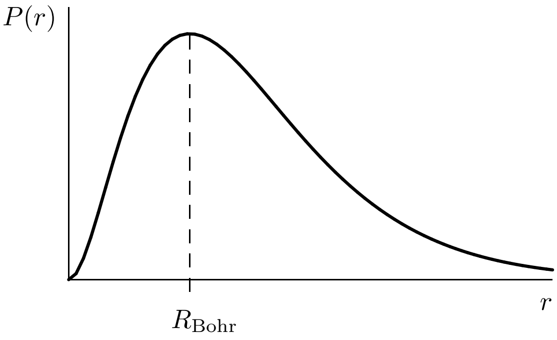


Figure 19.1: Probability density $P(r)$ for finding the electron at a distance r from the proton in the ground state of hydrogen atom.

at different distances from the nucleus would be different. What is amazing: the most probable distance corresponds to the classical trajectory!

You can visualize the electron inside an atom as moving around the nucleus chaotically and extremely fast so that for our "mental eyes" it forms a cloud. In some places this cloud is more dense while in other places more thin. The density of the cloud corresponds to the probability of finding the electron in a particular place. Space distribution of this density (probability) is what we can calculate using quantum mechanics. Results of such calculation for hydrogen atom are shown in Fig. 19.1. As was mentioned above, the most probable distance (maximum of the curve) coincides with the Bohr radius.

Quantum mechanical equation for any bound system (like an atom) can have solutions only at a discrete set of energies $E_1, E_2, E_3 \ldots$, etc. There are simply no solutions for the energies E in between these values, such as, for instance, $E_1 < E < E_2$. This is why a bound system of microscopic particles cannot have an arbitrary energy and can only be in one of the quantum states. Each of such states has certain energy and certain space configuration, i.e. distribution of the probability.

A bound quantum system can make transitions from one quantum state to another either spontaneously or as a result of interaction with other systems. The energy conservation law is one of the most fundamental and is valid in quantum world as well as in classical world. This means that any transition between the states with energies E_i and E_j is accompanied with either emission or absorption of the energy $\Delta E = |E_i - E_j|$. This is how an atom emits light.

Electron is a very light particle. Its mass is negligible as compared to the total mass of the atom. For example, in the lightest of all atoms, hydrogen, the electron constitutes only 0.054% of the atomic mass. In the silicon atoms that are the main component of the rocks around us, all 14 electrons make up only 0.027% of the mass. Thus, when holding a heavy rock in your hand, you actually feel the collective weight of all the nuclei that are inside it.

19.2 Nucleus

Is the nucleus a solid body? Is it an elementary building block of nature? No and no! Although it is very small, a nucleus consists of something even smaller.

19.2.1 Proton

The only way to do experiments with such small objects as atoms and nuclei, is to collide them with each other and watch what happens. Perhaps you think that this is a barbaric way, like colliding a "Mercedes" and "Toyota" in order to learn what is under their bonnets. But with microscopic particles nothing else can be done.

In the early 1920's Rutherford and other physicists made many experiments, changing one element into another by striking them with energetic helium nuclei. They noticed that all the time hydrogen nuclei were emitted in the process. It was apparent that the hydrogen nucleus played a fundamental role in nuclear structure and was a constituent part of all other nuclei. By the late 1920's physicists were regularly referring to hydrogen nucleus as *proton.* The term "proton" seems to have been coined by Rutherford, and first appears in print in 1920.

19.2.2 Neutron

Thus it was established that atomic nuclei consist of protons. Number of protons in a nucleus is such that makes up its positive charge. This number, therefore, coincides with the atomic number of the element in the Mendeleev's Periodic table.

This sounded nice and logical, but serious questions remained. Indeed, how can positively charged protons stay together in a nucleus? Repelling each other by electric force, they should fly away in different directions. Who keeps them together?

Furthermore, the proton mass is not enough to account for the nuclear masses. For example, if the protons were the only particles in the nucleus, then a helium nucleus (atomic number 2) would have two protons and therefore only twice the mass of hydrogen. However, it actually is four times heavier than hydrogen. This suggests that it must be something else inside nuclei in addition to protons.

These additional particles that kind of "glue" the protons and make up the nuclear mass, apparently, are electrically neutral. They were therefore called *neutrons.* Rutherford predicted the existence of the neutron in 1920. Twelve years later, in 1932, his assistant James Chadwick found it and measured its mass, which turned out to be almost the same but slightly larger than that of the proton.

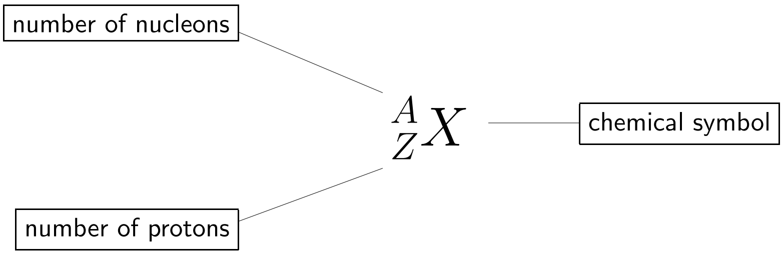
19.2.3 Isotopes

Thus, in the early 1930's it was finally proved that atomic nucleus consists of two types of particles, the protons and neutrons. The protons are positively charged while the neutrons are electrically neutral. The proton charge is exactly equal but opposite to that of the electron. The masses of proton and neutron are almost the same, approximately 1836 and 1839 electron masses, respectively.

Apart from the electric charge, the proton and neutron have almost the same properties. This is why there is a common name for them: *nucleon.* Both the proton and neutron are nucleons, like a man and a woman are both humans. In physics literature, the proton is denoted by letter p and the neutron by n. Sometimes, when the difference between them is unimportant, it is used the letter N meaning nucleon (in the same sense as using the word "person" instead of man or woman).

Chemical properties of an element are determined by the charge of its atomic nucleus, i.e. by the number of protons. This number is called the *atomic number* and is denoted by letter Z. The mass of an atom depends on how many nucleons its nucleus contains. The number of nucleons, i.e. total number of protons and neutrons, is called the *atomic mass number* and is denoted by letter A.

Standard nuclear notation shows the chemical symbol, the mass number and the atomic number of the isotope.



For example, the iron nucleus (26-th place in the Mendeleev's periodic table of the elements) with 26 protons and 30 neutrons is denoted as

$$^{56}_{26}\mathrm{Fe}\ ,$$

where the total nuclear charge is $Z = 26$ and the mass number $A = 56$. The number of neutrons is simply the difference $N = A - Z$ (here, it is used the same letter N, as for nucleon, but this should not cause any confusion). Chemical symbol is inseparably linked with Z. This is why the lower index is sometimes omitted and you may encounter the simplified notation like ^{56}Fe.

If we add or remove a few neutrons from a nucleus, the chemical properties of the atom remain the same because its charge is the same. This means that such atom should remain in the same place of the Periodic table. In Greek, "same place" reads *ίσος τόπος* (isos topos). The nuclei, having the same number of protons, but different number of neutrons, are called therefore *isotopes.*

Different isotopes of a given element have the same atomic number Z, but different mass numbers A since they have different numbers of neutrons N. Chemical properties of different isotopes of an element are identical, but they will often have great differences in nuclear stability. For stable isotopes of the light elements, the number of neutrons will be almost equal to the number of protons, but for heavier elements, the number of neutrons is always greater than Z and the neutron excess tends to grow when Z increases. This is because neutrons are kind of glue that keeps repelling protons together. The greater the repelling charge, the more glue you need.

19.3 Nuclear force

Since atomic nuclei are very stable, the protons and neutrons must be kept inside them by some force and this force must be rather strong. What is this force? All of modern particle physics was discovered in the effort to understand this force!

Trying to answer this question, at the beginning of the XX-th century, physicists found that all they knew before, was inadequate. Actually, by that time they knew only gravitational and electromagnetic forces. It was clear that the forces holding nucleons were not electromagnetic. Indeed, the protons, being positively charged, repel each other and all nuclei would decay in a split of a second if some other forces would not hold them together. On the other hand, it was also clear that they were not gravitational, which would be too weak for the task.

The simple conclusion was that nucleons are able to attract each other by yet unknown *nuclear forces*, which are stronger than the electromagnetic ones. Further studies proved that this hypothesis was correct.

Nuclear force has rather unusual properties. Firstly, it is charge independent. This means that in all pairs *nn*, *pp*, and *np* nuclear forces are the same. Secondly, at distances $\sim 10^{-13}$ cm, the nuclear force is attractive and very strong, $\sim$ 100 times stronger than the electromagnetic repulsion. Thirdly, the nuclear force is of a very short range. If the nucleons move away from each other for more than few *fermi* ($1\,\text{fm}=10^{-13}$ cm) the nuclear attraction practically disappears. Therefore the nuclear force looks like a "strong man with very short hands".

19.4 Binding energy and nuclear masses

19.4.1 Binding energy

When a system of particles is bound, you have to spend certain energy to disintegrate it, i.e. to separate the particles. The easiest way to do it is to strike the system with a moving particle that carries kinetic energy, like we can destroy a glass bottle with a bullet or a stone. If our bullet-particle moves too slow (i.e. does not have enough kinetic energy) it cannot disintegrate the system. On the other hand, if its kinetic energy is too high, the system is not only disintegrated but the separated particles acquire some kinetic energy, i.e. move away with some speed. There is an intermediate value of the energy which is just enough to destroy the system without giving its fragments any speed. This minimal energy needed to break up a bound system is called *binding energy* of this system. It is usually denoted by letter B.

19.4.2 Nuclear energy units

The standart unit of energy, *Joule*, is too large to measure the energies associated with individual nuclei. This is why in nuclear physics it is more convenient to use a much smaller unit called Mega-electron-Volt (MeV). This is the amount of energy that an electron acquires after passing between two charged plates with the potential difference (voltage) of one million Volts. Sounds very huge, isn't it? But look at this relation

$$1\,\text{MeV} = 1.602 \times 10^{-13}\,\text{J}$$

and think again. In the units of MeV, most of the energies in nuclear world can be expressed by values with only few digits before decimal point and without ten to the power of something. For example, the binding energy of proton and neutron (which is the simplest nuclear system and is called *deuteron*) is

$$B_{pn} = 2.225\,\text{MeV} \ .$$

The simplicity of the numbers is not the only advantage of using the unit MeV. Another, more important advantage, comes from the fact that most of experiments in nuclear physics are collision experiments, where particles are accelerated by electric field and collide with other particles. From the above value of B_{pn}, for instance, we immediately know that in order to break up deuterons, we need to bombard them with a flux of electrons accelerated through a voltage not less than 2.225 million Volts. No calculation is needed! On the other hand, if we know that a charged particle (with a unit charge) passes through a voltage, say, 5 million Volts, we can, without any calculation, say that it acqures the energy of 5 MeV. It is very convenient. Isn't it?

19.4.3 Mass defect

Comparing the masses of atomic nuclei with the masses of the nucleons that constitute them, we encounter a surprising fact: Total mass of the nucleons is greater than mass of the nucleus! For example, for the deuteron we have

$$m_d < m_p + m_n \ ,$$

where m_d, m_p, and m_n are the masses of deuteron, proton, and neutron, respectively. The difference is rather small,

$$(m_p + m_n) - m_d = 3.968 \times 10^{-30}\,\text{kg} \ ,$$

but on the nuclear scale is noticeable since the mass of proton, for example,

$$m_p = 1672.623 \times 10^{-30}\,\text{kg}$$

is also very small. This phenomenon is called *"mass defect"*. Where the mass disappears to, when nucleons are bound?

To answer this question, we notice that the energy of a bound state is lower than the energy of free particles. Indeed, to liberate them from a bound complex, we have to give them some energy. Thinking in the opposite direction, we conclude that, when forming a bound state, the particles have to get rid of the energy excess, which is exactly equal to the binding energy. This is observed experimentally: When a proton captures a neutron to form a deuteron, the excess energy of 2.225 MeV is emitted via electromagnetic radiation.

A logical conclusion from the above comes by itself: When proton and neutron are bounding, some part of their mass disappears together with the energy that is carried away by the radiation. And in the opposite process, when we break up the deutron, we give it the energy, some part of which makes up the lost mass.

Albert Einstein came to the idea of the equivalence between the mass and energy long before any experimental evidences were found. In his theory of relativity, he showed that total energy E of a moving body with mass m is

$$E = \frac{mc^2}{\sqrt{1 - \dfrac{v^2}{c^2}}} \ , \tag{19.1}$$

where v is its velocity and c the speed of light. Applying this equation to a non-moving body ($v = 0$), we conclude that it possesses the *rest* energy

$$E_0 = mc^2 \tag{19.2}$$

simply because it has mass. As you will see, this very formula is the basis for making nuclear bombs and nuclear power stations!

All the development of physics and chemistry, preceding the theory of relativity, was based on the assumption that the mass and energy of a closed system are conserving in all possible processes and they are conserved separately. In reality, it turned out that the conserving quantity is the *mass-energy*,

$$E_{\text{kin}} + E_{\text{pot}} + E_{\text{rad}} + mc^2 = \text{const} \ ,$$

i.e. the sum of kinetic energy, potential energy, the energy of radiation, and the mass of the system.

In chemical reactions the fraction of the mass that is transformed into other forms of energy (and vise versa), is so small that it is not detectable even in most precise measurements. In nuclear processes, however, the energy release is very often millions times higher and therefore is observable.

You should not think that mutual transformations of mass and energy are the features of only nuclear and atomic processes. If you break up a piece of rubber or chewing gum, for example, in two parts, then the sum of masses of these parts will be slightly larger than the mass of the whole piece. Of course we will not be able to detect this “mass defect” with our scales. But we can calculate it, using the Einstein formula (19.2). For this, we would need to measure somehow the mechanical work A used to break up the whole piece (i.e. the amount of energy supplied to it). This can be done by measuring the force and displacement in the breaking process. Then, according to Eq. (19.2), the mass defect is

$$\Delta m = \frac{A}{c^2} \ .$$

To estimate possible effect, let us assume that we need to stretch a piece of rubber in 10 cm before it breaks, and the average force needed for this is 10 N (approximately 1 kg). Then

$$A = 10\,\text{N} \times 0.1\,\text{m} = 1\,\text{J} \ ,$$

and hence

$$\Delta m = \frac{1\,\text{J}}{(299792458\,\text{m/s})^2} \approx 1.1 \times 10^{-17}\,\text{kg}.$$

This is very small value for measuring with a scale, but huge as compared to typical masses of atoms and nuclei.

19.4.4 Nuclear masses

Apparently, an individual nucleus cannot be put on a scale to measure its mass. Then how can nuclear masses be measured?

This is done with the help of the devices called *mass spectrometers*. In them, a flux of identical nuclei, accelerated to a certain energy, is directed to a screen where it makes a visible mark.

Before striking the screen, this flux passes through magnetic field, which is perpendicular to velocity of the nuclei. As a result, the flux is deflected to certain angle. The greater the mass, the smaller is the angle (because of inertia). Thus, measuring the displacement of the mark from the center of the screen, we can find the deflection angle and then calculate the mass.

Since mass and energy are equivalent, in nuclear physics it is customary to measure masses of all particles in the units of energy, namely, in MeV. Examples of masses of subatomic particles are given in Table 19.1. The values given in this table, are the energies to which the nuclear masses are equivalent via the Einstein formula (19.2).

particle	number of protons	number of neutrons	mass (MeV)
e	0	0	0.511
p	1	0	938.272
n	0	1	939.566
$^{2}_{1}\mathrm{H}$	1	1	1875.613
$^{3}_{1}\mathrm{H}$	1	2	2808.920
$^{3}_{2}\mathrm{He}$	2	1	2808.391
$^{4}_{2}\mathrm{He}$	2	2	3727.378
$^{7}_{3}\mathrm{Li}$	3	4	6533.832
$^{9}_{4}\mathrm{Be}$	4	5	8392.748
$^{12}_{6}\mathrm{C}$	6	6	11174.860
$^{16}_{8}\mathrm{O}$	8	6	14895.077
$^{238}_{92}\mathrm{U}$	92	146	221695.831

Table 19.1: Masses of electron, nucleons, and some nuclei.

There are several advantages of using the units of MeV to measure particle masses. First of all, like with nuclear energies, we avoid handling very small numbers that involve ten to the power of something. For example, if we were measuring masses in kg, the electron mass would be $m_e = 9.1093897 \times 10^{-31}$ kg. When masses are given in the equivalent energy units, it is very easy to calculate the mass defect. Indeed, adding the masses of proton and neutron, given in the second and third rows of Table 19.1, and subtracting the mass of $^{2}_{1}\mathrm{H}$, we obtain the binding energy 2.225 MeV of the deuteron without further ado. One more advantage comes from particle physics. In collisions of very fast moving particles new particles (like electrons) can be created from vacuum, i.e. kinetic energy is directly transformed into mass. If the mass is expressed in the energy units, we know how much energy is needed to create this or that particle, without calculations.

19.5 Radioactivity

As was said before, the nucleus experiences the intense struggle between the electric repulsion of protons and nuclear attraction of the nucleons to each other. It therefore should not be surprising that there are many nuclei that are unstable. They can spontaneously (i.e. without an external push) break in pieces. When the fragments reach the distances where the short range nuclear attraction disappears, they fiercely push each other away by the electric forces. Thus accelerated, they move in different directions like small bullets making destruction on their way. This is an example of nuclear radioactivity but there are several other varieties of radioactive decay.

19.5.1 Discovery of radioactivity

Nuclear radioactivity was discovered by Antoine Henri Becquerel in 1896. Following Wilhelm Roentgen who discovered the X-rays, Becquerel pursued his own investigations of these mysterious rays.

The material Becquerel chose to work with contained uranium. He found that the crystals containing uranium and exposed to sunlight, made images on photographic plates even wrapped in black paper. He mistakingly concluded that the sun's energy was being absorbed by the uranium which then emitted X-rays. The truth was revealed thanks to bad weather.

On the 26th and 27th of February 1896 the skies over Paris were overcast and the uranium crystals Becquerel intended to expose to the sun were returned to a drawer and put over (by chance) the photographic plates. On the first of March, Becquerel developed the plates and to his surprise, found that the images on them were clear and strong. Therefore the uranium emitted radiation without an external source of energy such as the sun. This was the first observation of the nuclear radioactivity.

Later, Becquerel demonstrated that the uranium radiation was similar to the X-rays but, unlike them, could be deflected by a magnetic field and therefore must consist of charged particles. For his discovery of radioactivity, Becquerel was awarded the 1903 Nobel Prize for physics.

19.5.2 Nuclear α, β, and γ rays

Classical experiment that revealed complex content of the nuclear radiation, was done as follows. The radium crystals (another radioactive element) were put at the bottom of a narrow straight channel made in a thick piece of lead and open at one side. The lead absorbed everything except the particles moving along the channel. This device therefore produced a flux of particles moving in one direction like bullets from a machine gun. In front of the channel was a photoplate that could register the particles.

Without the magnetic field, the image on the plate was in the form of one single dot. When the device was immersed into a perpendicular magnetic field, the flux of particles was split in three fluxes, which was reflected by three dots on the photographic plate.

One of the three fluxes was stright, while two others were deflected in opposite directions. This showed that the initial flux contained positive, negative, and neutral particles. They were named respectively the α, β, and γ particles.

The α-rays were found to be the ^{4}He nuclei, two protons and two neutrons bound together. They have weak penetrating ability, a few centimeters of air or a few sheets of paper can effectively block them. The β-rays proved to be electrons. They have a greater penetrating power than the α-particles and can penetrate 3 mm of aluminum. The γ-rays are not deflected because they are high energy photons. They have the same nature as the radio waves, visible light, and the X-rays, but have much shorter wavelength and therefore are much more energetic. Among the three, the γ-rays have the greatest penetrating power being able to pass through several centimeters of lead and still be detected on the other side.

19.5.3 Danger of the ionizing radiation

The α, β, and γ particles moving through matter, collide with atoms and knock out electrons from them, i.e. make positive ions out of the atoms. This is why these rays are called *ionizing radiation.*

Apart from ionizing the atoms, this radiation destroys molecules. For humans and all other organisms, this is the most dangerous feature of the radiation. Imagine thousands of tiny tiny bullets passing through your body and making destruction on their way. Although people do not feel any pain when exposed to nuclear radiation, it harms the cells of the body and thus can make people sick or even kill them. Illness can strike people years after their exposure to nuclear radiation. For example, the ionizing particles can randomly modify the DNA (long organic molecules that store all the information on how a particular cell should function in the body). As a result, some cells with wrong DNA may become cancer cells.

Fortunately, our body is able to repair some damages caused by radiation. Indeed, we are constantly bombarded by the radiation coming from the outer space as well as from the inner parts of our own planet and still survive. However, if the number of damages becomes too large, the body will not cope with them anymore.

There are established norms and acceptable limits for the radiation that are considered safe for human body. If you are going to work in contact with radioactive materials or near them, make sure that the exposure dose is monitored and the limits are adhered to.

You should understand that no costume can protect you from γ-rays! Only a thick wall of concrete or metal can stop them. The special costumes and masks that people wear when handling radioactive materials, protect them not from the rays but from contamination with that materials. Imagine if few specks of radioactive dirt stain your everyday clothes or if you inhale radioactive atoms. They will remain with you all the time and will shoot the "bullets" at you even when you are sleeping.

In many cases, a very effective way of protecting yourself from the radiation is to keep certain distance. Radiation from nuclear sources is distributed equally in all directions. Therefore the number n of dangerous particles passing every second through a unit area (say 1 cm^2) is the total number N of particles emitted during 1 second, divided by the surface of a sphere

$$n = \frac{N}{4\pi r^2} \ ,$$

where r is the distance at which we make the observation. From this simple formula, it is seen that the radiation intensity falls down with incresing distance quadratically. In other words, if you increase the distance by a factor of 2, your exposure to the radiation will be decreased by a factor of 4.

19.5.4 Decay law

Unstable nuclei decay spontaneously. A given nucleus can decay next moment, next day or even next century. Nobody can predict when it is going to happen. Despite this seemingly chaotic and "unscientific" situation, there is a strict order in all this.

Atomic nuclei, being microscopic objects, are ruled by quantum probabilistic laws. Although we cannot predict the exact moment of its decay, we can calculate the probability that a nucleus will decay within this or that time interval. Nuclei decay because of their internal dynamics and not because they become "old" or somehow "rotten".

To illustrate this, let us imagine that yesterday morning we found that a certain nucleus was going to decay within 24 hours with the probability of 50%. However, this morning we found that it is still "alive". This fact does not mean that the decay probability for another 24 hours increased. Not at all! It remains the same, 50%, because the nucleus remains the same, nothing wrong happened to it. This can go on and on for centuries.

Actually, we never deal with individual nuclei but rather with huge numbers of identical nuclei. For such collections (ensembles) of quantum objects, the probabilistic laws become statictical laws. Let us assume that in the above example we had 1 million identical nuclei instead of only one. Then by this morning only half of these nuclei would survive because the decay probability for 24 hours was 50%. Among the remaining 500000 nuclei, 250000 will decay by tomorrow morning, then after another 24 hours only 125000 will remain and so on.

The number of unstable nuclei that are still "alive" continuously decreases with time according to the curve shown in Fig. 19.2. If initially, at time $t = 0$, their number is N_0, then after certain time interval $T_{1/2}$ only half of these nuclei will remain, namely, $\frac{1}{2}N_0$. Another one half of the remaining half will decay during another such interval. So, after the time $2T_{1/2}$, we will have only one quarter of the initial amount, and so on. The time interval $T_{1/2}$, during which one half of unstable nuclei decay, is called their *half-life time.* It is specific for each unstable nucleus and vary from a fraction of a second to thousands and millions of years. A few examples of such lifetimes are given in Table 19.2

19.5.5 Radioactive dating

Examining the amounts of the decay products makes possible radioactive dating. The most famous is the *Carbon dating*, a variety of radioactive dating which is applicable only to matter which was once living and presumed to be in equilibrium with the atmosphere, taking in carbon dioxide from the air for photosynthesis.

Cosmic ray protons blast nuclei in the upper atmosphere, producing neutrons which in turn bombard nitrogen, the major constituent of the atmosphere. This neutron bombardment produces the radioactive isotope ${}^{14}_{6}$C. The radioactive carbon-14 combines with oxygen to form

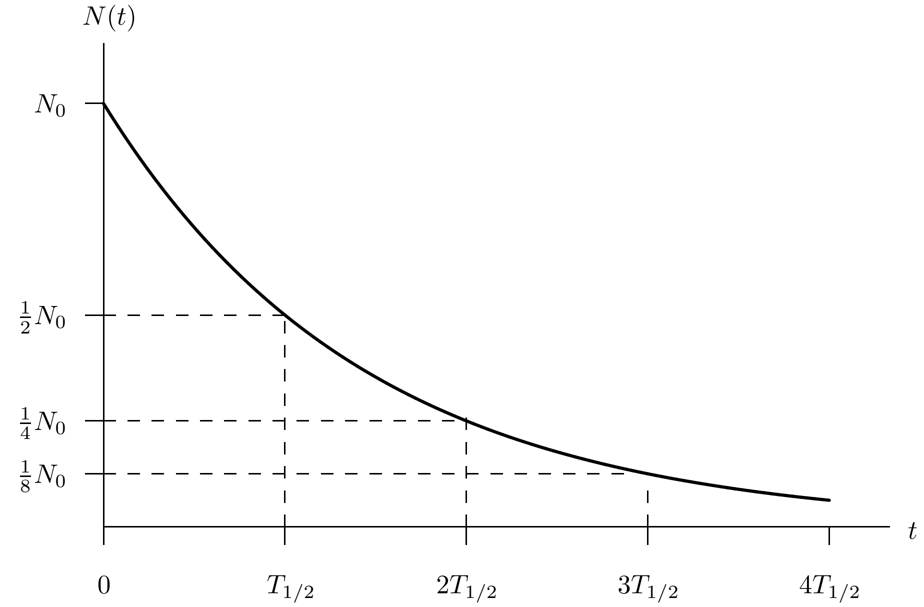


Figure 19.2: The time $T_{1/2}$ during which one half of the initial amount of unstable particles decay, is called their half-life time.

isotope	$T_{1/2}$	decay mode
$^{214}_{84}\mathrm{Po}$	$1.64 \times 10^{-4}\,\mathrm{s}$	α, γ
$^{89}_{36}\mathrm{Kr}$	$3.16\,\mathrm{min}$	β^-, γ
$^{222}_{86}\mathrm{Rn}$	$3.83\,\mathrm{days}$	α, γ
$^{90}_{38}\mathrm{Sr}$	$28.5\,\mathrm{years}$	β^-
$^{226}_{88}\mathrm{Ra}$	$1.6 \times 10^{3}\,\mathrm{years}$	α, γ
$^{14}_{6}\mathrm{C}$	$5.73 \times 10^{3}\,\mathrm{years}$	β^-
$^{238}_{92}\mathrm{U}$	$4.47 \times 10^{9}\,\mathrm{years}$	α, γ
$^{115}_{49}\mathrm{In}$	$4.41 \times 10^{14}\,\mathrm{years}$	β^-

Table 19.2: Half-life times of several unstable isotopes.

carbon dioxide and is incorporated into the cycle of living things.

The isotope $^{14}_{6}\mathrm{C}$ decays (see Table 19.2) inside living bodies but is replenished from the air and food. Therefore, while an organism is alive, the concentration of this isotope in the body remains constant. After death, the replenishment from the breath and food stops, but the isotopes that are in the dead body continue to decay. As a result the concentration of ${}^{14}_{6}\mathrm{C}$ in it gradually decreases according to the curve shown in Fig. 19.2. The time $t = 0$ on this Figure corresponds to the moment of death, and N_0 is the equilibrium concentration of ${}^{14}_{6}\mathrm{C}$ in living organisms.

Therefore, by measuring the radioactive emissions from once-living matter and comparing its activity with the equilibrium level of emissions from things living today, an estimation of the time elapsed can be made. For example, if the rate of the radioactive emissions from a piece of wood, caused by the decay of ${}^{14}_{6}\mathrm{C}$, is one-half lower than from living trees, then we can conclude that we are at the point $t = T_{1/2}$ on the curve 19.2, i.e. it is elapsed exactly one half-life-time period. According to the Table 19.2), this means that the tree, from which this piece of wood was made, was cut approximately 5730 years ago. This is how physicists help archaeologists to assign dates to various organic materials.

19.6 Nuclear reactions

Those of you who studied chemistry, are familiar with the notion of chemical reaction, which, in essence, is just regrouping of atoms that constitute molecules. As a result, reagent chemical compounds are transformed into product compounds.

In the world of nuclear particles, similar processes are possible. When nuclei are close to each other, nucleons from one nucleus can "jump" into another one. This happens because there are attractive and repulsive forces between the nucleons. The complicated interplay of these forces may cause their regrouping. As a result, the reagent particles are transformed into product particles. Such processes are called *nuclear reactions.*

For example, when two isotopes ${}^{3}_{2}\mathrm{He}$ collide, the six nucleons constituting them, can rearrange in such a way that the isotope ${}^{4}_{2}\mathrm{He}$ is formed and two protons are liberated. Similarly to chemical reactions, this process is denoted as

$${}^{3}_{2}\mathrm{He} + {}^{3}_{2}\mathrm{He} \longrightarrow {}^{4}_{2}\mathrm{He} + p + p + 12.86\,\mathrm{MeV}\ . \tag{19.3}$$

The same as in chemical reactions, nuclear reactions can also be either exothermic (i.e. releasing energy) or endothermic (i.e. requiring an energy input). The above reaction releases 12.86 MeV of energy. This is because the total mass on the left hand side of Eq. (19.3) is in 12.86 MeV greater than the total mass of the products on the right hand side (you can check this using Table 19.1).

Thus, when considering a particular nuclear reaction, we can always learn if it releases or absorbs energy. For this, we only need to compare total masses on the left and right hand sides of the equation. Now, you can understand why it is very convenient to express masses in the units of energy.

Composing equations like (19.3), we should always check the superscripts and subscripts of the nuclei in order to have the same number of nucleons and the same charge on both sides of the equation. In the above example, we have six nucleons and the charge +4 in both the initial and final states of the reaction. To make the checking of nucleon number and charge conservation easier, sometimes the proton and neutron are denoted with superscripts and subscripts as well, namely, ${}_1^1p$ and ${}_0^1n$. In this case, all we need is to check that sum of superscripts and sum of subscripts are the same on both sides of the equation.

19.7 Detectors

How can we observe such tiny tiny things as protons and α-particles? There is no microscope that would be able to discern them. From the very beginning of the sub-atomic era, scientists have been working on the development of special instruments that are called particle detectors. These devices enable us either to register the mere fact that certain particle has passed through certain point in space or to observe the trace of its path (the trajectory). Actually, this is as good as watching the particle. Although the particle sizes are awfully small, when passing through some substances, they leave behind visible traces of tens of centimeters in length. By measuring the curvature of the trajectory of a particle deflected in electric or magnetic field, a physicist can determine the charge and mass of the particle and thus can identify it.

19.7.1 Geiger counter

The most familiar device for registering charged particles is the Geiger counter. It cannot tell you anything about the particle except the fact that it has passed through the counter. The counter consists of a thin metal cylinder filled with gas. A wire electrode runs along the center of the tube and is kept at a high voltage ($\sim 2000\,\mathrm{V}$) relative to the cylinder. When a particle passes through the tube, it causes ionization of the gas atoms and thus an electric discharge between the cylinder and the wire. The electric pulse can be counted by a computer or made to produce a "click" in a loudspeaker. The number of counts per second tells us about intensity of the radiation.

19.7.2 Fluorescent screen

The very first detector was the fluorescent screen. When a charged particle hits the screen, a human eye can discern a flash of light at the point of impact. In fact, we all use this kind of detectors every day when watching TV of looking at a computer (if it does not have an LCD screen of course). Indeed, the images on the screens of their electron-ray tubes are formed by the accelerated electrons.

19.7.3 Photo-emulsion

Another type of particle detector, dating back to Becquerel, is the nuclear photographic emulsion. Passage of charged particles is recorded in the emulsion in the same way that ordinary black and white photographic film records a picture. The only difference is that nuclear photoemulsion is made rather thick in order to catch a significant part of the particle path. After the developing, a permanent record of the charged particle trajectory is available.

19.7.4 Wilson's chamber

In the fields of sub-atomic physics and nuclear physics, Wilson's cloud chamber is the most fundamental device to observe the trajectories of particles. Its basic principle was discovered by C. T. R. Wilson in 1897, and it was put to the practical use in 1911.

The top and the side of the chamber are covered with round glasses of several centimeters in diameter. At the bottom of the chamber, a piston is placed. The air filled in the chamber is saturated with vapor of water. When pulling down the piston quickly, the volume of the chamber would be expanded and the temperature goes down. As a result, the air inside would be supersaturated with the vapor. If a fast moving charged particle enters the chamber when it is in such a supersaturated state, the vapor of water would condense along the line of the ions generated by the particle, which is the path of the particle. Thus we can observe the trace, and also take a photograph. To make clear the trace, a light is sometimes illuminated from the side. When placing the cloud chamber in a magnetic field, we can obtain various informations about the charged particle by measuring the curvature of the trace and other data. The bubble chamber and the spark chamber have taken place of the cloud chamber which is nowadays used only for the educational purposes. Wilson's cloud chamber has however played a very important role in the history of physics.

19.7.5 Bubble chamber

Bubble chamber is a particle detector of major importance during the initial years of high-energy physics. The bubble chamber has produced a wealth of physics from about 1955 well into the 1970s. It is based on the principle of bubble formation in a liquid heated above its boiling point, which is then suddenly expanded, starting boiling where passing charged particles have ionized the atoms of the liquid. The technique was honoured by the Nobel prize award to D. Glaser in 1960. Even today, bubble chamber photographs provide the aesthetically most appealing visualization of subnuclear collisions.

19.7.6 Spark chamber

Spark chamber is a historic device using electric discharges over a gap between two electrodes with large potential difference, to render passing particles visible. Sparks occurred where the gas had been ionized. Most often, multiple short gaps were used, but wide-gap chambers with gaps up to 40 cm were also built. The spark chamber is still of great scientific value in that it remains relatively simple and cheap to build as well as enabling an observer to view the paths of charged particles.

19.8 Nuclear energy

Nuclei can produce energy via two different types of reactions, namely, *fission* and *fusion* reactions. Fission is a break up of a nucleus in two or more pieces (smaller nuclei). Fusion is the opposite process: Formation of a bigger nucleus from two small nuclei.

A question may arise: How two opposite processes can both produce energy? Can we make an inexhaustible souce of energy by breaking up and then fusing the same nuclei? Of cousre not! The energy conservation law cannot be circumvented in no way. When speaking about fusion and fission, we speak about different ranges of nuclei. Energy can only be released when either light nuclei fuse or heavy nuclei fission.

To understand why this is so, let us recollect that for releasing energy the mass of initial nuclei must be greater than the mass of the products of a nuclear reaction. The mass difference is transformed into the released energy. And why the product nuclei can loose some mass as compared to the initial nuclei? Because they are more tightly bound, i.e. their binding energies are lager.

Fig. 19.3 shows the dependence of the binding energy B per nucleon on the number A of nucleons constituting a nucleus. As you see, the curve reaches the maximum value of $\sim 9\,\mathrm{MeV}$ per nucleon at around $A \sim 50$. The nuclei with such number of nucleons cannot produce energy neither through fusion nor through fission. They are kind of “ashes” and cannot serve as a fuel. In contrast to them, very light nuclei, when fused with each other, make more tightly bound products as well as very heavy nuclei do when split up in lighter fragments.

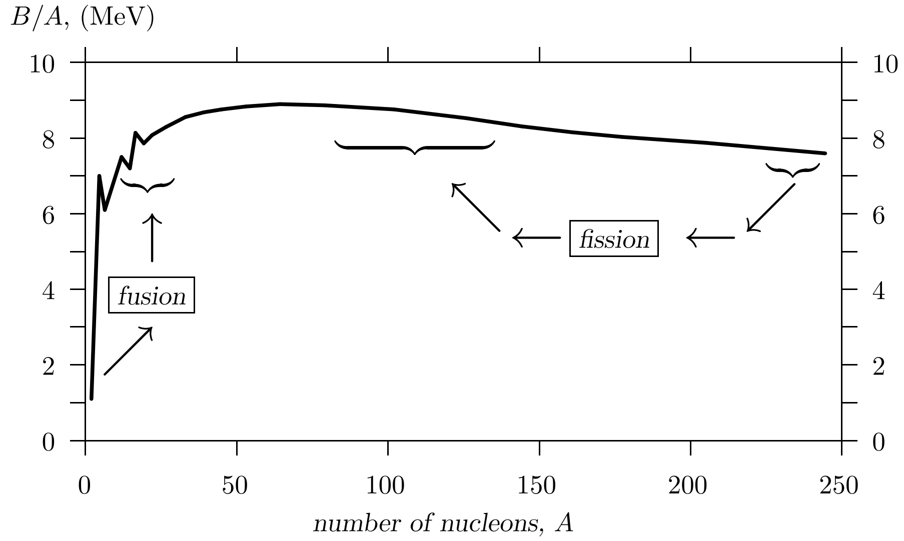


Figure 19.3: Binding energy per nucleon.

In fission processes, which were discovered and used first, a heavy nucleus like, for example, uranium or plutonium, splits up in two fragments which are both positively charged. These fragments repel each other by an electric force and move apart at a high speed, distributing their kinetic energy in the surrounding material.

In fusion reactions everything goes in the opposite direction. Very light nuclei, like hydrogen or helium isotopes, when approaching each other to a distance of a few fm ($1\,\mathrm{fm} = 10^{-13}\,\mathrm{cm}$), experience strong attraction which overpowers their Coulomb (that is electric) repulsion. As a result the two nuclei fuse into a single nucleus. They collapse with extremely high speeds towards each other. To form a stable nucleus they must get rid of the excessive energy. This energy is emitted by ejecting a neutron or a photon.

19.8.1 Nuclear reactors

Since the discovery of radioactivity it was known that heavy nuclei release energy in the processes of spontaneous decay. This process, however, is rather slow and cannot be influenced (speed up or slow down) by humans and therefore could not be effectively used for large-scale energy production. Nonetheless, it is ideal for feeding the devices that must work autonomously in remote places for a long time and do not require much energy. For this, heat from the spontaneous-decays can be converted into electric power in a radioisotope thermoelectric generator. These generators have been used to power space probes and some lighthouses built by Russian engineers. Much more effective way of using nuclear energy is based on another type of nuclear decay which is considered next.

Chain reaction

The discovery that opened up the era of nuclear energy was made in 1939 by German physicists O. Hahn, L. Meitner, F Strassmann, and O. Frisch. They found that a uranium nucleus, after absorbing a neutron, splits into two fragments. This was not a spontaneous but induced fission

$$n + {}^{235}_{92}\mathrm{U} \longrightarrow {}^{140}_{54}\mathrm{Xe} + {}^{94}_{38}\mathrm{Sr} + n + n + 185\,\mathrm{MeV} \qquad (19.4)$$

that released $\sim 185\,\mathrm{MeV}$ of energy as well as two neutrons which could cause similar reactions on surrounding nuclei. The fact that instead of one initial neutron, in the reaction (19.4) we obtain two neutrons, is crucial. This gives us the possibility to make the so-called *chain reaction* schematically shown in Fig. 19.4.

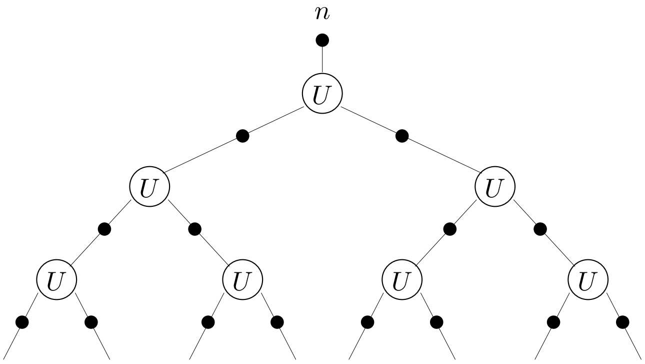


Figure 19.4: Chain reaction on uranium nuclei.

In such process, one neutron breaks one heavy nucleus, the two released neutrons break two more heavy nuclei and produce four neutrons which, in turn, can break another four nuclei and so on. This process develops extremely fast. In a split of a second a huge amount of energy can be released, which means explosion. In fact, this is how the so-called atomic bomb works.

Can we control the development of the chain reaction? Yes we can! This is done in nuclear reactors that produce energy for our use. How can it be done?

Critical mass

First of all, if the piece of material containing fissile nuclei is too small, some neutrons may reach its surface and escape without causing further fissions. For each type of fissile material there is therefore a minimal mass of a sample that can support explosive chain reaction. It is called the *critical mass.* For example, the critical mass of ${}^{235}_{92}\mathrm{U}$ is approximately 50 kg. If the mass is below the critical value, nuclear explosion is not possible, but the energy is still released and the sample becomes hot. The closer mass is to its critical value, the more energy is released and more intensive is the neutron radiation from the sample.

The criticality of a sample (i.e. its closeness to the critical state) can be reduced by changing its geometry (making its surface bigger) or by putting inside it some other material (boron or cadmium) that is able to absorb neutrons. On the other hand, the criticality can be increased by putting neutron reflectors around the sample. These reflectors work like mirrors from which the escaped neutrons bounce back into the sample. Thus, moving in and out the absorbing material and reflectors, we can keep the sample close to the critical state.

How a nuclear reactor works

In a typical nuclear reactor, the fuel is not in one piece, but in the form of several hundred vertical rods, like a brush. Another system of rods that contain a neutron absorbing material (control rods) can move up and down in between the fuel rods. When totally in, the control rods absorb so many neutrons, that the reactor is shut down. To start the reactor, operator gradually moves the control rods up. In an emergency situation they are dropped down automatically.

To collect the energy, water flows through the reactor core. It becomes extremely hot and goes to a steam generator. There, the heat passes to water in a secondary circuit that becomes steam for use outside the reactor enclosure for rotating turbines that generate electricity.

Nuclear power in South Africa

By 2004 South Africa had only one commercial nuclear reactor supplying power into the national grid. It works in Koeberg located 30 km north of Cape Town. A small research reactor was also operated at Pelindaba as part of the nuclear weapons program, but was dismantled.

Koeberg Nuclear Power station is a uranium Pressurized Water Reactor (PWR). In such a reactor, the primary coolant loop is pressurised so the water does not boil, and heat exchangers, called steam generators, are used to transmit heat to a secondary coolant which is allowed to boil to produce steam. To remove as much heat as possible, the water temperature in the primary loop is allowed to rise up to about 300 °C which requires the pressure of 150 atmospheres (to keep water from boiling).

The Koeberg power station has the largest turbine generators in the southern hemisphere and produces ∼10000 MWh of electric energy. Construction of Koeberg began in 1976 and two of its Units were commissioned in 1984-1985. Since then, the plant has been in more or less continuous operation and there have been no serious incidents.

Eskom that operates this power station, may be the current technology leader. It is developing a new type of nuclear reactor, a modular pebble-bed reactor (PBMR). In contrast to traditional nuclear reactors, in this new type of reactors the fuel is not assembled in the form of rods. The uranium, thorium or plutonium fuels are in oxides (ceramic form) contained within spherical pebbles made of pyrolitic graphite. The pebbles, having a size of a tennis ball, are in a bin or can. An inert gas, helium, nitrogen or carbon dioxide, circulates through the spaces between the fuel pebbles. This carries heat away from the reactor.

Ideally, the heated gas is run directly through a turbine. However since the gas from the primary coolant can be made radioactive by the neutrons in the reactor, usually it is brought to a heat exchanger, where it heats another gas, or steam.

The primary advantage of pebble-bed reactors is that they can be designed to be inherently safe. When a pebble-bed reactor gets hotter, the more rapid motion of the atoms in the fuel increases the probability of neutron capture by $^{238}_{92}$U isotopes through an effect known as Doppler broadening. This isotope does not split up after capturing a neutron. This reduces the number of neutrons available to cause $^{235}_{92}$U fission, reducing the power output by the reactor. This natural negative feedback places an inherent upper limit on the temperature of the fuel without any operator intervention.

The reactor is cooled by an inert, fireproof gas, so it cannot have a steam explosion as a water reactor can.

A pebble-bed reactor thus can have all of its supporting machinery fail, and the reactor will not crack, melt, explode or spew hazardous wastes. It simply goes up to a designed "idle" temperature, and stays there. In that state, the reactor vessel radiates heat, but the vessel and fuel spheres remain intact and undamaged. The machinery can be repaired or the fuel can be removed.

A large advantage of the pebble bed reactor over a conventional water reactor is that they operate at higher temperatures. The reactor can directly heat fluids for low pressure gas turbines. The high temperatures permit systems to get more mechanical energy from the same amount of thermal energy.

Another advantage is that fuel pebbles for different fuels might be used in the same basic design of reactor (though perhaps not at the same time). Proponents claim that some kinds of pebble-bed reactors should be able to use thorium, plutonium and natural unenriched Uranium, as well as the customary enriched uranium. One of the projects in progress is to develop pebbles and reactors that use the plutonium from surplus or expired nuclear explosives.

On June 25, 2003, the South African Republic's Department of Environmental Affairs and Tourism approved ESKOM's prototype 110 MW pebble-bed modular reactor for Koeberg. Eskom also has approval for a pebble-bed fuel production plant in Pelindaba. The uranium for this fuel is to be imported from Russia. If the trial is successful, Eskom says it will build up to ten local PBMR plants on South Africa's seacoast. Eskom also wants to export up to 20 PBMR plants per year. The estimated export revenue is 8 billion rand a year, and could employ about 57000 people.

19.8.2 Fusion energy

For a given mass of fuel, a fusion reaction like

$$^2_1\mathrm{H} + ^3_1\mathrm{H} \longrightarrow ^4_2\mathrm{He} + n + 17.59\,\mathrm{MeV}\ . \tag{19.5}$$

yield several times more energy than a fission reaction. This is clear from the curve given in Fig. 19.3. Indeed, a change of the binding energy (per nucleon) is much more significant for a fusion reaction than for a fission reaction. Fusion is, therefore, a much more powerful source of energy. For example, 10 g of Deuterium which can be extracted from 500 litres of water and 15 g of Tritium produced from 30 g of Lithium would give enough fuel for the lifetime electricity needs of an average person in an industrialised country.

But this is not the only reason why fusion attracted so much attention from physicists. Another, more fundamental, reason is that the fusion reactions were responsible for the synthesis of the initial amount of light elements at primordial times when the universe was created. Furthermore, the synthesis of nuclei continues inside the stars where the fusion reactions produce all the energy which reaches us in the form of light.

Thermonuclear reactions

If fusion is so advantageous, why is it not used instead of fission reactors? The problem is in the electric repulsion of the nuclei. Before the nuclei on the left hand side of Eq. (19.5) can fuse, we have to bring them somehow close to each other to a distance of $\sim 10^{-13}$ cm. This is not an easy task! They both are positively charged and "refuse" to approach each other.

What we can do is to make a mixture of the atoms containing such nuclei and heat it up. At high temperatures the atoms move very fast. They fiercely collide and loose all the electrons. The mixture becomes *plasma*, i.e. a mixture of bare nuclei and free moving electrons. If the temperature is high enough, the colliding nuclei can overcome the electric repulsion and approach each other to a fusion distance.

When the nuclei fuse, they release much more energy than was spent to heat up the plasma. Thus the initial energy "investment" pays off. The typical temperature needed to ignite the reaction of the type (19.5) is extremely high. In fact, it is the same temperature that our sun has in its center, namely, $\sim$15 million degrees. This is why the reactions (19.3), (19.5), and the like are called *thermonuclear reactions.*

Human-made thermonuclear reactions

The same as with fission reactions, the first application of thermonuclear reactions was in weapons, namely, in the hydrogen bomb, where fusion is ignited by the explosion of an ordinary (fission) plutonium bomb which heats up the fuel to solar temperatures.

In an attempt to make a controllable fusion, people encounter the problem of holding the plasma. It is relatively easy to achieve a high temperature (with laser pulses, for example). But as soon as plasma touches the walls of the container, it immediately cools down. To keep it from touching the walls, various ingenious methods are tried, such as strong magnetic field and laser beams directed to plasma from all sides. In spite of all efforts and ingenious tricks, all such attempts till now have failed. Most probably this straightforward approach to controllable fusion is doomed because one has to hold in hands a "piece of burning sun".

Cold fusion

To visualize the struggle of the nuclei approaching each other, imagine yourself pushing a metallic ball towards the top of a slope shown in Fig. 19.5. The more kinetic energy you give to the ball, the higher it can climb. Your purpose is to make it fall into the narrow well that is behind the barrier.

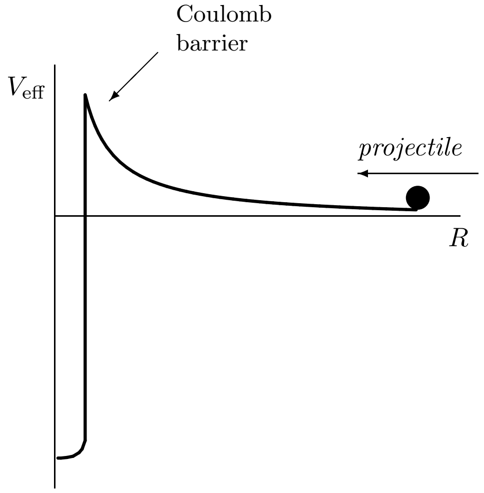


Figure 19.5: Effective nucleus–nucleus potential as a function of the separation between the nuclei.

In fact, the curve in Fig. 19.5 shows the dependence of relative potential energy $V_{\rm eff}$ between two nuclei on the distance R separating them. The deep narrow well corresponds to the strong short-range attraction, and the $\sim 1/R$ barrier represents the Coulomb (electric) repulsion. The nuclei need to overcome this barrier in order to "touch" each other and fuse, i.e. to fall into the narrow and deep potential well. One way to achieve this is to give them enough kinetic energy, which means to rise the temperature. However, there is another way based on the quantum laws.

As you remember, when discussing the motion of the electron inside an atom (see Sec. 19.1), we said that it formed a "cloud" of probability around the nucleus. The density of this cloud diminishes at very short and very long distances but never disappears completely. This means that we can find the electron even inside the nucleus though with a rather small probability.

The nuclei moving towards each other, being microscopic objects, obey the quantum laws as well. The probability density for finding one nucleus at a distance R from another one also forms a cloud. This density is non-zero even under the barrier and on the other side of the barrier. This means that, in contrast to classical objects, quantum particles, like nuclei, can penetrate through potential barriers even if they do not have enough energy to go over it! This is called the *tunneling effect.*

The tunneling probability strongly depends on thickness of the barrier. Therefore, instead of lifting the nuclei against the barrier (which means rising the temperature), we can try to make the barrier itself thinner or to keep them close to the barrier for such a long time that even a low penetration probability would be realized.

How can this be done? The idea is to put the nuclei we want to fuse, inside a molecule where they can stay close to each other for a long time. Furthermore, in a molecule, the Coulomb barrier becomes thinner because of electron screening. In this way fusion may proceed even at room temperature.

This idea of *cold fusion* was originally (in 1947) discussed by F. C. Frank and (in 1948) put forward by A. D. Sakharov, the "father" of Russian hydrogen bomb, who at the latest stages of his career was worldwide known as a prominent human rights activist and a winner of the Nobel Prize for Peace. When working on the bomb project, he initiated research into peaceful applications of nuclear energy and suggested the fusion of two hydrogen isotopes via the reaction (19.5) by forming a molecule of them where one of the electrons is replaced by a muon.

The muon is an elementary particle (see Sec. 19.9), which has the same characteristics as an electron. The only difference between them is that the muon is 200 times heavier than the electron. In other words, a muon is a heavy electron. What will happen if we make a muonic atom of hydrogen, that is a bound state of a proton and a muon? Due to its large mass the muon would be very close to the proton and the size of such atom would be 200 times smaller than that of an ordinary atom. This is clearly seen from the formula for the atomic Bohr radius

$$R_{\text{Bohr}} = \frac{\hbar^2}{me^2} ,$$

where the mass is in the denominator.

Now, what happens if we make a muonic molecule? It will also be 200 times smaller than an ordinary molecule. The Coulomb barrier will be 200 times thinner and the nuclei 200 times closer to each other. This is just what we need! Speaking in terms of the effective nucleus–nucleus potential shown in Fig. 19.5, we can say that the muon modifies this potential in such a way that a second minimum appears. Such a modified potential is (schematically) shown in Fig. 19.6.

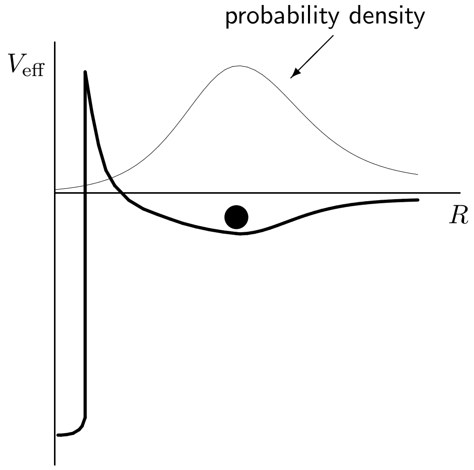


Figure 19.6: Effective nucleus–nucleus potential (thick curve) for nuclei confined in a molecule. Thin curve shows the corresponding distribution of the probability for finding the nuclei at a given distance from each other.

The molecule is a bound state in the shallow but wide minimum of this potential. Most of the time, the nuclei are at the distance corresponding to the maximum of the probability density distribution (shown by the thin curve). Observe that this density is not zero under the barrier (though is rather small) and even at $R = 0$. This means that the system can (with a small probability) jump from the shallow well into the deep well through the barrier, i.e. can tunnel and fuse.

Unfortunately, the muon is not a stable particle. Its lifetime is only $\sim 10^{-6}$ sec. This means that a muonic molecule cannot exist longer than 1 microsecond. As a matter of fact, from a quantum mechanical point of view, this is quite a long interval.

The quantum mechanical wave function (that describes the probability density) oscillates with a frequency which is proportional to the energy of the system. With a typical binding energy of a muonic molecule of 300 eV this frequency is $\sim 10^{17}\,\mathrm{s}^{-1}$. This means that the particle hits the barrier with this frequency and during 1 microsecond it makes 10^{11} attempts to jump through it. The calculations show that the penetration probability is $\sim 10^{-7}$. Therefore, during 1 microsecond nuclei can penetrate through the barrier 10000 times and fusion can happen much faster than the decay rate of the muon.

Cold fusion via the formation of muonic molecules was done in many laboratories, but unfortunately, it cannot solve the problem of energy production for our needs. The obstacle is the negative efficiency, *i.e.* to make muonic cold fusion we have to spend more energy than it produces. The reason is that muons do not exist like protons or electrons. We have to produce them in accelerators. This takes a lot of energy.

Actually, the muon serves as a catalyst for the fusion reaction. After helping one pair of nuclei to fuse, the muon is liberated from the molecule and can form another molecule, and so on. It was estimated that the efficiency of the energy production would be positive only if each muon ignited at least 1000 fusion events. Experimentalists tried their best, but by now the record number is only 150 fusion events per muon. This is too few. The main reason why the muon does not catalyze more reactions is that it is eventually trapped by a ^{4}He nucleus which is a by-product of fusion. Helium captures the muon into an atomic orbit with large binding energy, and it cannot escape.

Nonetheless, the research in the field of cold fusion continues. There are some other ideas of how to keep nuclei close to each other. One of them is to put the nuclei inside a crystal. Another way out is to increase the penetration probability by using molecules with special properties, namely, those that have quantum states with almost the same energies as the excited states on the compound nucleus. Scientists try all possibilities since the energy demands of mankind grow continuously and therefore the stakes in this quest are high.

19.9 Elementary particles

In our quest for the elementary building blocks of the universe, we delved inside atomic nucleus and found that it is composed of protons and neutrons. Are the three particles, e, p, and n, the blocks we are looking for? The answer is “no”. Even before the structure of the atom was understood, Becquerel discovered the redioactivity (see Sec. 19.5.1) that afterwards puzzled physicists and forced them to look deeper, i.e. inside protons and neutrons.

19.9.1 β decay

Among the three types of radioactivity, the α and γ rays were easily explained. The emission of α particle is kind of fission reaction, when an initial nucleus spontaneously decays in two fragments one of which is the nucleus ${}^4_2\mathrm{He}$ (i.e. α particle). The γ rays are just electromagnetic quanta emitted by a nuclear system when it transits from one quantum state to another (the same like an atom emits light).

The β rays posed the puzzle. On the one hand, they are just electrons and you may think that it looks simple. But on the other hand, they are not the electrons from the atomic shell. It was found that they come from inside the nucleus! After the β-decay, the charge of the nucleus increases in one unit,

$$ {}^{A}_{Z}\,(\textsf{parent nucleus}) \longrightarrow {}_{Z+1}^{\;\;A}\,(\textsf{daughter nucleus}) + e \ , $$

which is in accordance with the charge conservation law.

There was another puzzle associated with the β decay: The emitted electrons did not have a certain energy. Measuring their kinetic energies, you could find very fast and very slow electrons as well as the electrons with all intermediate speeds. How could identical parent nuclei, after loosing different amount of energy, become identical daughter nuclei. May be energy is not conserving in the quantum world? The fact was so astonishing that even Niels Bohr put forward the idea of statistical nature of the energy conservation law.

To explain the first puzzle, it was naively suggested that neutron is a bound state of proton and electron. At that time, physicists believed that if something is emitted from an object, this something must be present inside that object before the emission. They could not imagine that a particle could be created from vacuum.

The naive (pe) model of the neutron contradicted the facts. Indeed, it was known already that the pe bound state is the hydrogen atom. Neutron is much smaller than the atom. Therefore, it would be unusually tight binding, and perhaps with something elese involved that keeps the size small. By the way, this "something elese" could also save the energy conservation law. In 1930, Wolfgang Pauli suggested that in addition to the electron, the β decay involves another particle, ν, that is emitted along with the electron and carries away part of the energy. For example,

$$ {}^{234}_{\ 90}\mathrm{Th} \longrightarrow {}^{234}_{\ 91}\mathrm{Pa} + e^- + \bar{\nu} \ . \tag{19.6} $$

This additional particle was called *neutrino* (in Italian the word "neutrino" means small neutron). The neutrino is electrically neutral, has extremely small mass (maybe even zero, which is still a question in 2004) and very weakly interacts with matter. This is why it was not detected experimentally till 1956. The "bar" over ν in Eq. (19.6) means that in this reaction actually the anti-neutrino is emitted (see the discussion on anti-particles further down in Sec. 19.9.2).

19.9.2 Particle physics

In an attempt to explain the β decay and to understand internal structure of the neutron a new branch of physics was born, the *particle physics.* The only way to explore the structure of subatomic particles is to strike them with other particles in order to knock out their "constituent" parts. The simple logic says: The more powerful the impact, the smaller parts can be knocked out.

At the beginning the only source of energetic particles to strike other particles were the cosmic rays. Earth is constantly bombarded by all sort of particles coming from the outer space. Atmosphere protects us from most of them, but many still reach the ground.

Antiparticles

In 1932, studying the cosmic rays with a bubble chamber, Carl Anderson made a photograph of two symmetrical tracks of charged particles. The measurements of the track curvatures showed that one track belonged to an electron and the other was made by a particle having the same mass and equal but positive charge. These particles were created when a cosmic γ quantum of a high energy collided with a nucleus.

The discovered particle was called *positron* and denoted as e^+ to distinguish it from the electron, which sometimes is denoted as e^-. It was the first antiparticle discovered. Later, it was found that every particle has its "mirror reflection", the antiparticle. To denote an antiparticle, it is used "bar" over a particle symbol. For example, $\bar{p}$ is the anti-proton, which has the same mass as an ordinary proton but a negative charge.

When a particle collides with its "mirror reflection", they annihilate, i.e. they burn out completely. In this collision, all their mass is transformed into electromagnetic energy in the form of γ quanta. For example, if an electron collides with a positron, the following reaction may take place

$$e^- + e^+ \longrightarrow \gamma + \gamma \ , \tag{19.7}$$

where two photons are needed to conserve the total momentum of the system.

In principle, stable antimatter can exist. For example, the pair of $\bar{p}$ and e^+ can form an atom of anti-hydrogen with exactly the same energy states as the ordinary hydrogen. Experimentally, atoms of anti-helium were obtained. The problem with them is that, surrounded by ordinary matter, they cannot live long. Colliding with ordinary atoms, they annihilate very fast.

There are speculations that our universe should be symmetric with respect to particles and antiparticles. Indeed, why should preference be given to matter and not to anti-matter? This implies that somewhere very far, there must be equal amount of anti-matter, i.e. anti-universe. Can you imagine what happens if they meet?

Muon, mesons, and the others

In yet another cosmic-ray experiment a particle having the same properties as the electron but $\sim$207 times heavier, was discovered in 1935. It was given the name *muon* and the symbol μ. For a long time it remained "unnecessary" particle in the picture of the world. Only the modern theories harmonically included the muon as a constituent part of matter (see Sec 19.9.3).

The same inexhaustible cosmic rays revealed the π and K mesons in 1947. The π mesons (or simply *pions*) were theoretically predicted twelve years before by Yukawa, as the mediators of the strong forces between nucleons. The K mesons, however, were unexpected. Furthermore, they showed very strange behaviour. They were easily created only in pairs. The probability of the inverse process (i.e. their decay) was 10^{13} times lower than the probability of their creation.

It was suggested that these particles possess a new type of charge, the *strangeness*, which is conserving in the strong interactions. When a pair of such particles is created, one of them has strangeness $+1$ and the other -1, so the total strangeness remains zero. When decaying, they act individually and therefore the strangeness is not conserving. According to the suggestion, this is only possible through the weak interactions that are much weaker than the strong interactions (see Sec. 19.9.4) and thus the decay probability is much lower.

The golden age of particle physics began in 1950-s with the advent of particle accelerators, the machines that produced beams of electrons or protons with high kinetic energy. Having such beams available, experimentalists can plan the experiment and repeat it, while with the cosmic rays they were at the mercy of chance. When the accelerators became the main tool of exploration, the particle physics acquired its second name, the *high energy physics*.

During the last half a century, experimentalists discovered so many new particles (few of them are listed in Table 19.3) that it became obvious that they cannot all be elementary. When colliding with each other, they produce some other particles. Mutual transformations of the particles is their main property.

family	particle	symbol	mass (MeV)	Lifetime $T_{1/2}$ (s)
photon	photon	γ	0	stable
leptons	electron	e^-, e^+	0.511	stable
	muon	μ^-, μ^+	105.7	2.2×10^{-6}
	tau	τ^-, τ^+	1777	10^{-13}
	electron neutrino	ν_e	~ 0	stable
	muon neutrino	ν_μ	~ 0	stable
	tau neutrino	ν_τ	~ 0	stable
hadrons	pion	π^+, π^-	139.6	2.6×10^{-8}
	pion	π^0	135.0	0.8×10^{-16}
	kaon	K^+, K^-	493.7	1.2×10^{-8}
	kaon	K^0_S	497.7	0.9×10^{-10}
	kaon	K^0_L	497.7	5.2×10^{-8}
	eta meson	η^0	548.8	10^{-18}
	proton	p	938.3	stable
	neutron	n	939.6	900
	lambda	Λ^0	1116	2.6×10^{-10}
	sigma	Σ^+	1189	0.8×10^{-10}
	sigma	Σ^0	1192	6×10^{-20}
	sigma	Σ^-	1197	1.5×10^{-10}
	omega	Ω^-, Ω^+	1672	0.8×10^{-10}

Table 19.3: Few representatives of different particle families.

Physicists faced the problem of particle classification similar to the problems of classification of animals, plants, and chemical elements. The first approach was very simple. The particles were divided in four groups according to their mass: *leptons* (light particles, like electron), *mesons* (intermediate mass, like pion), *baryons* (heavy particles, like proton or neutron), and *hyperons* (very heavy particles).

Then it was realized that it would be more logical to divide the particles in three families according to their ability to interact via weak, electromagnetic, and strong forces (in addition to that, all particles experience gravitational attraction towards each other). Except for the gravitational interaction, the photon (γ quantum) participates only in electromagnetic interactions, the leptons take part in both weak and electromagnetic interactions, and *hadrons* are able to interact via all forces of nature (see Sec. 19.9.4).

In addition to conservation of the strangeness, several other conservation laws were discovered. For example, number of leptons is conserving. This is why in the reaction (19.6) we have an electron (lepton number $+1$) and anti-neutrino (lepton number -1) in the final state. Similarly, the number of baryons is conserving in all reactions.

The quest for the constituent parts of the neutron has led us to something unexpected. We found that there are several hundreds of different particles that can be "knocked out" of the neutron but none of them are its parts. Actually, the neutron itself can be knocked out of some of them! What a mess! Further efforts of experimentalists could not find an order, which was finally discovered by theoreticians who introduced the notion of *quarks*.

19.9.3 Quarks and leptons

While experimentalists seemed to be lost in the maze, the theoreticians groped for the way out. Using an extremely complicated mathematical technique, they managed to group the hadrons in such families which implied that all known (and yet unknown) hadrons are build of only six types of particles with fractional charges. The main credit for this (in the form of Nobel Prize) was given to M. Gell-Mann and G. Zweig.

At first, they considered a subset of the hadrons and developed a theory with only three types of such truly elementary particles. When Murray Gell-Mann thought of the name for them, he came across the book "Finnegan's Wake" by James Joyce. The line "Three quarks for Mister Mark..." appeared in that fanciful book (in German, the word "quark" means cottage cheese). He needed a name for three particles and this was the answer. Thus the term *quark* was coined.

Later, the theory was generalized to include all known particles, which required six types of quarks. Modern theories require also that the number of different leptons should be the same as the number of different quark types. According to these theories, the quarks and leptons are truly elementary, i.e. they do not have any internal structure and therefore are of a zero size (pointlike). Thus, the world is constructed of just twelve types of elementary building blocks that are given in Table 19.4. Amazingly enough, the electron that was discovered before all other particles, more than a century ago, turned out to be one of them!

After Gell-Mann, who used a funny name (quark) for an elementary particle, the fundamental physics was flooded with such names. For example, the six quark types are called *flavors* (for cottage cheese, this is appropriate indeed), the three different states in which each quark can be, are called *colors* (red, green, blue), etc. Modern physics is so complicated and mathematical, that people working in it, need such kind of jokes to "spice unsavoury dish with flavors". The funny names should not confuse anybody. Elementary particles do not have any smell, taste, or colour. These terms simply denote certain properties (similar to electric charge) that do not

family	elementary particle	symbol	charge	lepton number	baryon number	mass (MeV)
leptons	electron	e^-	-1	1	0	0.511
	muon	μ^-	-1	1	0	105.7
	tau	τ^-	-1	1	0	1777
	electron neutrino	ν_e	0	1	0	~ 0
	muon neutrino	ν_μ	0	1	0	~ 0
	tau neutrino	ν_τ	0	1	0	~ 0
quarks	up	u	$+2/3$	0	$1/3$	360
	down	d	$-1/3$	0	$1/3$	360
	strange	s	$-1/3$	0	$1/3$	1500
	charmed	c	$+2/3$	0	$1/3$	540
	top (truth)	t	$+2/3$	0	$1/3$	174000
	bottom (beauty)	b	$-1/3$	0	$1/3$	5000

Table 19.4: Elementary building blocks of the universe.

exist in human world.

Hadrons

There are particles that are able to interact with each other by the so-called *strong forces.* Another name for these forces is *nuclear forces.* They are very strong at short distances ($\sim 10^{-15}$ m), and very quickly vanish when the distance between the particles increases. All these particles are called *hadrons.* The protons and neutrons are examples of hadrons.

As you remember, we learned about the existence of huge variety of particles when trying to look inside a nucleon, more particularly, the neutron. So, what the neutron is made of? Can we get the answer at last, after learning about the quarks? Yes, we can.

According to modern theories, all hadrons are composed of quarks. The quarks can be combined in groups of two or three. The bound states of two quarks are called *mesons*, and the bound complexes of three quarks are called *baryons.* No other numbers of quarks can form observable particles[1].

Nucleons are baryons and therefore consist of three quarks while the pion is a meson containing only two quarks, as schematically shown in Fig. 19.7. Comparing this figure with Table 19.4, you can see why quarks have fractional charges. Counting the total charge of a hadron, you should not forget that anti-quarks have the opposite charges. The baryon number for an anti-quark also has the opposite sign (negative). This is why mesons actually consist of a quark and anti-quark in order to have total baryon number zero.

[1] Recently, experimentalists and theoreticians started to actively discuss the possibility of the existence of *pentaquarks*, exotic particles that are bound complexes of five quarks.

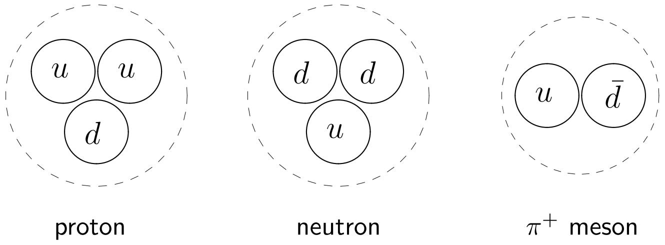


Figure 19.7: Quark content of the proton, neutron, and π^+-meson.

Particle reactions

At the early stages of the particle physics development, in order to find the constituent parts of various particles, experimentalists simply collided them and watched the "fragments". However, this straightforward approach led to confusion. For example, the reaction between the π^- meson and proton,

$$\pi^- + p \longrightarrow K^0 + \Lambda^0 \ , \tag{19.8}$$

would suggest (if naively interpreted) that either K^0 or Λ^0 is a constituent part of the nucleon while the pion is incorporated into the other "fragment". On the other hand, the same collision can knock out different "fragments" from the same proton. For example,

$$\pi^- + p \longrightarrow \pi^0 + n \ , \tag{19.9}$$

which leads to an absurd suggestion that neutron is a constituent part of proton.

The quark model explains all such "puzzles" nicely and logically. Similarly to chemical reactions that are just rearrangements of atoms, the particle reactions of the type (19.8) and (19.9) are just rearrangements of the quarks. The only difference is that, in contrast to chemistry where the number of atoms is not changing, the number of quarks before the collision is not necessarily equal to their number after the collision. This is because a quark from one colliding particle can annihilate with the corresponding antiquark from another particle. Moreover, if the collision is sufficiently powerful, the quark-antiquark pairs can be created from vacuum.

It is convenient to depict the particle transformations in the form of the so-called *quark flow diagrams*. On such diagrams, the quarks are represented by lines that may be visualized as the trajectories showing their movement from the left to the right.

For example, the diagram given in Fig. (19.8), shows the quark rearrangement for the reaction (19.8). As you can see, when the pion collides with proton, its $\bar{u}$ quark annihilates with the u quark from the proton. At the same time, the $s\bar{s}$ pair is created from the vacuum. Then, the $\bar{s}$ quark binds with the d quark to form the strange meson K^0, while the s quark goes together with the ud pair as the strange baryon Λ^0.

The charge-exchange reaction (19.9) is a more simple rearrangement process shown in Fig. 19.9. You may wonder why the quark and antiquark of the same flavor in the π^0 meson do not annihilate. Yes they do, but not immediately. And due to this annihilation, the lifetime of π^0 is 100 million times shorter than the lifetime of $\pi^\pm$ (see Table 19.3).

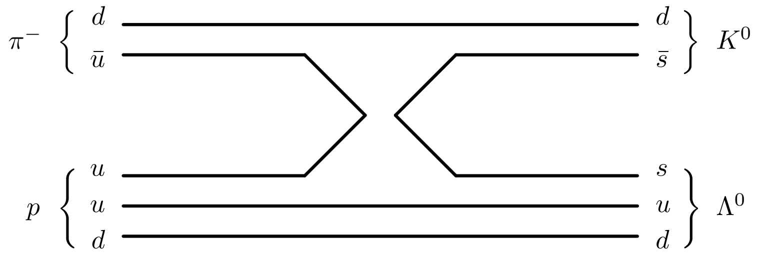


Figure 19.8: Quark-flow diagram for the reaction $\pi^- + p \longrightarrow K^0 + \Lambda^0$.

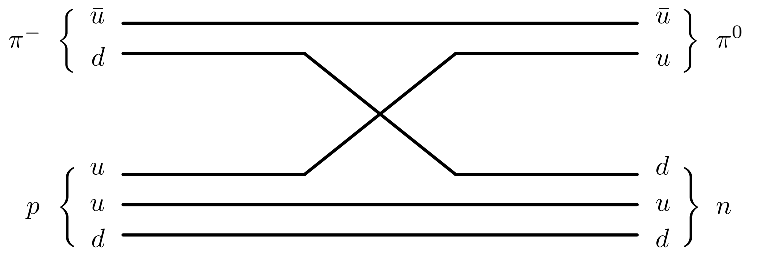


Figure 19.9: Quark-flow diagram for the reaction $\pi^- + p \longrightarrow \pi^0 + n$.

Despite its simplicity, the quark-flow diagram technique is very powerful method not only for explaining the observed reactions but also for predicting new reactions that have not yet been seen in experiments. Knowing the quark content of particles (which is available in modern Physics Handbooks), you can draw plenty of such diagrams that will describe possible particle transformations. The only rule is to keep the lines continuous. They can disappear or emerge only for a quark-antiquark pair of the same flavor.

However, the continuity of the quark lines is valid only for the processes caused by the strong interaction. Indeed, the β-decay of a free neutron (caused by the weak forces),

$$n \longrightarrow p + e^- + \bar{\nu}_e \ , \tag{19.10}$$

as well as the β-decay of the nuclei, indicate that quarks can change flavor. In particular, the β-decay (19.10) or (19.6) happens because the d quark transformes into the u quark,

$$d \longrightarrow u + e^- + \bar{\nu}_e \ , \tag{19.11}$$

due to the weak interaction, as shown in Fig. 19.10

Quark confinement

At this point, it is very logical to ask if anybody observed an isolated quark. The answer is "no". Why? And how can one be so confident of the quark model when no one has ever seen an

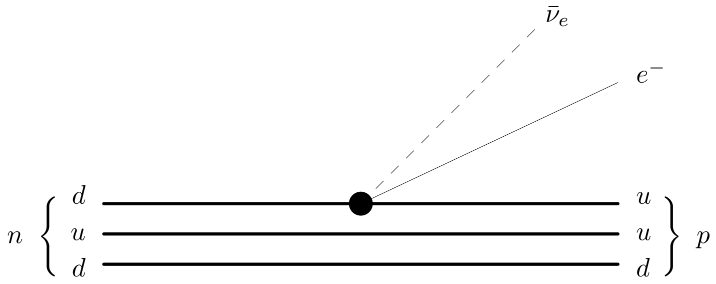


Figure 19.10: Quark-flow diagram for the β decay of neutron.

isolated quark?

Basically, you can't see an isolated quark because the quark-quark attractive force does not let them go. In contrast to all other systems, the attraction between quarks grows with the distance separating them. It is like a rubber cord connecting two balls. When the balls are close to each other, the cord is not stretched and the balls do not feel any force. If, however, you try to separate the balls, the cord pulls them back. The more you stretch the cord, the stronger the force becomes (according to the Hook's law of elasticity). Of course, a real rubber cord would eventually break. This does not happen with the quark-quark force. It can grow to infinity. This phenomenon is called the *confinement of quarks*.

Nonetheless, we are sure that the nucleon consists of three quarks having fractional charges. A hundred years ago Rutherford, by observing the scattering of charged particles from an atom, proved that its positive charge is concentrated in a small nucleus. Nowadays, similar experiments prove the existence of fractional point-like charges inside the nucleon.

The quark model actually is much more complicated than the quark-flow diagrams. It is a consistent mathematical theory that explains a vast variety of experimental data. This is why nobody doubts that it reflects the reality.

19.9.4 Forces of nature

If asked how many types of forces exist, many people start counting on their fingers, and when the count exceeds ten, they answer "plenty of". Indeed, there are gravitational forces , electrical, magnetic, elastic, frictional forces, and also forces of wind, of expanding steam, of contracting muscles, etc.

If, however, we analyze the root causes of all these forces, we can reduce their number to just a few fundamental forces (or *fundamental interactions*, as physicists say).

For example, the elastic force of a stretched rubber cord is due to the attraction between the molecules that the rubber is made of. Looking deeper, we find that the molecules attract each other because of the electromagnetic attraction between the electrons of one molecule and nuclei of the other. Similarly, if we depress a piece of rubber, it resists because the molecules refuse to approach each other too close due to the electric repulsion of the nuclei. Therefore the elasticity of rubber has the electromagnetic origin.

Any other force in the human world can be analyzed in the same manner. After doing this, we will find that all forces that we see around us (in the macroworld), are either of gravitational or electromagnetic nature. As we also know, in the microworld there are two other types of forces: The strong (nuclear) forces that act between all hadrons, and the weak forces that are responsible for changing the quark flavors.

Therefore, all interactions in the Universe are governed by only four fundamental forces: Strong, electromagnetic, weak and gravitational. These forces are very different in strength and range. Their relative strengths are given in Table 19.5. The most strong is the nuclear interaction. The strength of the electromagnetic forces is one hundred times lower. The weak forces are nine orders of magnitude weaker than the nuclear forces, and the gravity is 38 orders of magnitude weaker! It is amazing that this subtle interaction governs the cosmic processes. The reason is that the gravitational forces are of long range and always attractive. There is no such thing as negative mass that would screen the gravitational field, like negative electrons screen the field of positive nuclei.

Force	Relative Strength	Range
Strong	1	Short
Electromagnetic	0.0073	Long
Weak	10^{-9}	Very Short
Gravitational	10^{-38}	Long

Table 19.5: Four fundamental forces and their relative strengths.

Towards the unified force

Physicists always try to simplify things. Since there are only four fundamental forces, it is tempting to ask "If only four, then why not only one?". Can it be that all interactions are just different faces of one master force?

The first who started the quest for unification of forces was Einstein. After completing his general theory of relativity, he spent 30 years in unsuccessful attempts to unify the electromagnetic and gravity forces. At that time, it seemed logical because both of them were infinite in range and obeyed the same inverse square law. Einstein failed because the unification should be done on the basis of quantum laws, but he tried to do it using the classical concepts.

Electro-weak unification

Now it is known that despite the similarities in form of the gravity and electromagnetic forces, the gravity will be the last to yield to unification. The more implausible unification of the electromagnetic and weak forces turned out to be the first successful step towards the unified interaction.

In 1979, the Nobel prize was awarded to Weinberg, Salam, and Glashow, who developed a unified theory of electromagnetic and weak interactions. According to that theory, the electromagnetic and weak forces converge to one *electro-weak* interaction at very high collision energies. The theory also predicted the existence of heavy particles, the W and Z, with masses around 80000 MeV and 90000 MeV, respectively. These particles were discovered in 1983, which brought experimental verification to the new theory.

Grand unification

The next step was to try to combine the electro-weak theory with the theory of the strong interactions (i.e. quark theory) in a single theory. This work was called the *grand unification*. Currently, physicists discuss versions of such theory that predicts the convergence of the three forces at awfully high energies $\sim 10^{17}$ MeV. The quarks and leptons in this theory, are the unified *leptoquarks*.

The grand unification is not that successful as the electro-weak theory. It has the problem of mathematical consistency and contradicts to at least one experiment. The matter is that it predicts the proton decay,

$$p \longrightarrow e^+ + \pi^0 \ ,$$

that does not conserve both the baryon and lepton numbers, with the lifetime of $\sim 10^{29}$ years. The measurements show, however, that the lifetime of the proton is at least 10^{32} years.

Theory of everything

Some people believe that the grand unification has an inherent principal flaw. According to them, one cannot unify the forces step by step (leaving the gravity out), and the correct way is to combine all four forces in the so-called *theory of everything*.

There are few different approaches to unifying everything. One of them suggests that all fundamental particles (quarks and leptons) are just vibrating modes of string loops in multidimensional space. The electron is a string vibrating one way, the up-quark is a string vibrating another way, and so on. The other approach introduces a new level of fundamental particles, the *preons*, that could be constituent parts of quarks and leptons. The quest goes on.

Everyone agrees that constructing the theory of everything would in no way mean that biology, geology, chemistry, or even physics had been solved. The universe is so rich and complex that the discovery of the fundamental theory would not mean the end of science. The ultimate theory of everything would provide an unshakable pillar of coherence forever assuring us that the universe is a comprehensible place.

19.10 Origin of the universe

Looking deep inside microscopic particles, physicists need to collide them with high kinetic energies. The smaller parts of matter they want to observe, the higher energy they need. This is why they build more and more powerful accelerators. However, the accelerators have natural limitations. Indeed, an accelerator cannot be bigger than the size of our planet. And even if we manage to build a circular accelerator around the whole earth (along the equator, for example), it would not be able to reach the energy of $\sim 10^{17}$ MeV at which the grand unification of fundamental interactions takes place.

So, what are we to do? How can we test the theory of everything? Is it possible at all? Yes, it is! The astronomically high values, like $\sim 10^{17}$ MeV, should be looked for in the cosmos, of course. Our journey towards extremely small objects eventually leads us to extremely large objects, like whole universe.

Equations of Einstein's theory of relativity can describe the evolution of the universe. Physicists solved these equations back in time and found that the universe had its beginning. Approximately 15 billion years ago, it started from a zero size point that exploded and rapidly expanded to the present tremendous scale. At the first instants after the explosion, the matter was at such incredibly high density and temperature that all particles had kinetic energies even higher than the unification energy $\sim 10^{17}$ MeV. This means that at the very beginning there was only one sigle force and no difference among fundamental particles. Everything was unified and "simple".

You may ask "So what? How can so distant past help us?". In many ways! The development of the universe was governed by the fundamental forces. If our theories about them are correct, we should be able to reproduce (with calculations) how that development proceeded step by step. During the expansion, all the nuclei and atoms in the cosmos were created. The amounts of different nuclei are not the same. Why? Their relative abundances were determined by the processes in the first moments after the explosion. Thus, comparing what follows from the theories with the observed abundances of chemical elements, we can judge validity of our theories.

Nowadays, the most popular theory, describing the history of the universe, is the so–called *Big-Bang* model. The diagram given in Fig. 19.11, shows the sequence of events which led to the creation of matter in its present form.

Nobody knows what was before the Big Bang and why it happened, but it is assumed that just after this enigmatic cataclysm, the universe was so dense and hot that all four forces of nature (strong, electromagnetic, weak, and gravitational) were indistinguishable and therefore gravity was governed by quantum laws, like the other three types of interactions. A complete theory of quantum gravity has not been constructed yet, and this very first "epoch" of our history remains as enigmatic as the Big Bang itself.

The ideal "democracy" (equality) among the forces lasted only a small fraction of a second. By the time $t \sim 10^{-43}$ sec the universe cooled down to $\sim 10^{32}$ K and the gravity separated. The other three forces, however, remained unified into one universal interaction mediated by an extremely heavy particle, the so-called X boson, which could transform leptons into quarks and vice versa.

When at $t \sim 10^{-35}$ sec most of the X bosons decayed, the quarks combined in trios and pairs to form nucleons, mesons, and other hadrons. The only symmetry which lasted up to $\sim 10^{-10}$ sec, was between the electromagnetic and weak forces mediated by the Z and W particles. From the moment when this last symmetry was broken ($\sim 10^{-10}$ sec) until the universe was about one second old, neutrinos played the most significant role by mediating the neutron-proton transmutations and therefore fixing their balance (neutron to proton ratio).

Already in a few seconds after the Big Bang nuclear reactions started to occur. The protons

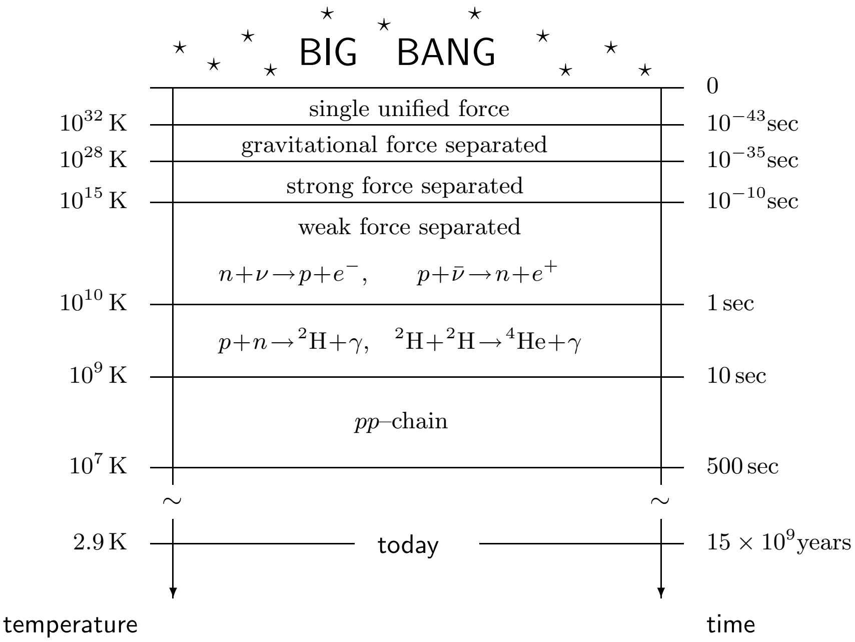


Figure 19.11: Schematic "history" of the universe.

and neutrons combined very rapidly to form deuterium and then helium. During the very first seconds there were too many very energetic photons around which destroyed these nuclei immediately after their formation. Very soon, however, the continuing expansion of the universe changed the conditions in favour of these newly born nuclei. The density decreased and the photons could not destroy them that fast anymore.

During a short period of cosmic history, between about 10 and 500 seconds, the entire universe behaved as a giant nuclear fusion reactor burning hydrogen. This burning took place via a chain of nuclear reactions, which is called the *pp-chain* because the first reaction in this sequence is the proton-proton collision leading to the formation of a deuteron. Nowadays, the same *pp*-chain is the main source of energy in our sun and other stars.

But how do we know that the scenario was like this? In other words, how can we check the Big–Bang theory? Is it possible to prove something which happened 15 billion years ago and in such a short time? Yes, it is! The *pp*-chain fusion,

pp-chain:

$$\boxed{\begin{array}{rcl} \mathrm{p}+\mathrm{p} & \to & {}^2\mathrm{H}+\mathrm{e}^+ +\nu_e \\ \mathrm{e}^- {+}\mathrm{p}+\mathrm{p} & \to & {}^2\mathrm{H}+\nu_e \\ \mathrm{p}+{}^2\mathrm{H} & \to & {}^3\mathrm{He}+\gamma \\ {}^3\mathrm{He}+{}^3\mathrm{He} & \to & {}^4\mathrm{He}+\mathrm{p}+\mathrm{p} \\ {}^3\mathrm{He}+{}^4\mathrm{He} & \to & {}^7\mathrm{Be}+\gamma \end{array}}$$

$$\boxed{\begin{aligned} e^- + {}^7\mathrm{Be} &\to {}^7\mathrm{Li} + \nu_e \\ p + {}^7\mathrm{Li} &\to {}^8\mathrm{Be} + \gamma \\ {}^8\mathrm{Be} &\to {}^4\mathrm{He} + {}^4\mathrm{He} \end{aligned}} \quad \boxed{\begin{aligned} p + {}^7\mathrm{Be} &\to {}^8\mathrm{B} + \gamma \\ {}^8\mathrm{B} &\to {}^8\mathrm{Be}^* + e^+ + \nu_e \\ {}^8\mathrm{Be}^* &\to {}^4\mathrm{He} + {}^4\mathrm{He} \end{aligned}}$$

is the key for such a proof.

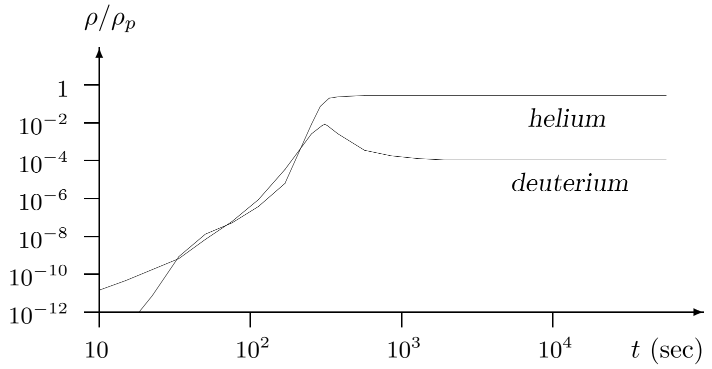


Figure 19.12: Mass fractions ρ (relative to hydrogen ρ_p) of primordial deuterium and ^{4}He versus the time elapsed since the Big Bang.

As soon as the nucleosynthesis started, the amount of deuterons, helium isotopes, and other light nuclei started to increase. This is shown in Fig. 19.12 for ^{2}H and ^{4}He. The temperature and the density, however, continued to decrease. After a few minutes the temperature dropped to such a level that the fusion practically stopped because the kinetic energy of the nuclei was not sufficient to overcome the electric repulsion between nuclei anymore. Therefore the abundances of light elements in the cosmos were fixed (we call them the *primordial abundances*). Since then, they practically remain unchanged, like a photograph of the past events, and astronomers can measure them. Comparing the measurements with the predictions of the theory, we can check whether our assumptions about the first seconds of the universe are correct or not.

Astronomy and the physics of microworld come to the same point from different directions. The Big Bang theory is only one example of their common interest. Another example is related to the mass of neutrino. When Pauli suggested this tiny particle to explain the nuclear β-decay, it was considered as massless, like the photon. However, the experiments conducted recently, indicate that neutrinos may have small non-zero masses of just a few eV.

In the world of elementary particles, this is extremely small mass, but it makes a huge difference in the cosmos. The universe continues to expand despite the fact that the gravitational forces pull everything back to each other. The estimates show, that the visible mass of all galaxies is not sufficient to stop and reverse the expansion. The universe is filled with a tremendous number of neutrinos. Even with few eV per neutrino, this amounts to a huge total mass of them, which is invisible but could reverse the expansion.

Thus, the cooperation of astronomers and particle physicists has led to significant advances in our understanding of the universe and its evolution. The quest goes on. A famous German philosopher Friedrich Nietzsche once said that "The most incomprehensible thing about this Universe is that it is comprehensible."

Appendix A

GNU Free Documentation License

Version 1.2, November 2002

PREAMBLE

The purpose of this License is to make a manual, textbook, or other functional and useful document “free” in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or non-commercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of “copyleft”, which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

APPLICABILITY AND DEFINITIONS

This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The “Document”, below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as “you”. You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A “Modified Version” of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A “Secondary Section” is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document’s overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The “Invariant Sections” are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The “Cover Texts” are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A “Transparent” copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not “Transparent” is called “Opaque”.

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The “Title Page” means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, “Title Page” means the text near the most prominent appearance of the work’s title, preceding the beginning of the body of the text.

A section “Entitled XYZ” means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as “Acknowledgements”, “Dedications”, “Endorsements”, or “History”.) To “Preserve the Title” of such a section when you modify the Document means that it remains a section “Entitled XYZ” according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

VERBATIM COPYING

You may copy and distribute the Document in any medium, either commercially or non-commercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section A.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

COPYING IN QUANTITY

If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

MODIFICATIONS

You may copy and distribute a Modified Version of the Document under the conditions of sections A and A above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

1. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History

section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.

2. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
3. State on the Title page the name of the publisher of the Modified Version, as the publisher.
4. Preserve all the copyright notices of the Document.
5. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
6. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
7. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
8. Include an unaltered copy of this License.
9. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
10. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
11. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
12. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
13. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
14. Do not re-title any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
15. Preserve any Warranty Disclaimers.

If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of

Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled “Endorsements”, provided it contains nothing but endorsements of your Modified Version by various parties–for example, statements of peer review or that the text has been approved by an organisation as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

COMBINING DOCUMENTS

You may combine the Document with other documents released under this License, under the terms defined in section A above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled “History” in the various original documents, forming one section Entitled “History”; likewise combine any sections Entitled “Acknowledgements”, and any sections Entitled “Dedications”. You must delete all sections Entitled “Endorsements”.

COLLECTIONS OF DOCUMENTS

You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

AGGREGATION WITH INDEPENDENT WORKS

A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an “aggregate” if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section A is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

TRANSLATION

Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section A. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section A) to Preserve its Title (section A) will typically require changing the actual title.

TERMINATION

You may not copy, modify, sub-license, or distribute the Document except as expressly provided for under this License. Any other attempt to copy, modify, sub-license or distribute the Document is void, and will automatically terminate your rights under this License. However, parties who have received copies, or rights, from you under this License will not have their licenses terminated so long as such parties remain in full compliance.

FUTURE REVISIONS OF THIS LICENSE

The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See http://www.gnu.org/copyleft/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation.

ADDENDUM: How to use this License for your documents

To use this License in a document you have written, include a copy of the License in the document and put the following copyright and license notices just after the title page:

> Copyright © YEAR YOUR NAME. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.2 or any later version published by the Free Software Foundation; with no Invariant Sections, no Front-Cover Texts, and no Back-Cover Texts. A copy of the license is included in the section entitled "GNU Free Documentation License".

If you have Invariant Sections, Front-Cover Texts and Back-Cover Texts, replace the "with...Texts." line with this:

with the Invariant Sections being LIST THEIR TITLES, with the Front-Cover Texts being LIST, and with the Back-Cover Texts being LIST.

If you have Invariant Sections without Cover Texts, or some other combination of the three, merge those two alternatives to suit the situation.

If your document contains nontrivial examples of program code, we recommend releasing these examples in parallel under your choice of free software license, such as the GNU General Public License, to permit their use in free software.