Modern Physics

Part One The Birth of a New Physics

Classical physics has its origins in the 17th century, with Galileo's' experiments on falling bodies, Kepler's calculations of the planetary orbits and Newton's postulates and mathematical laws of motion.

During the following two centuries, the Newtonian system of mechanics, together with its theory of gravitation, dominated scientific thinking; its achievements were unprecedented.

Perhaps its most impressive feat was predicting the existence of the planet Neptune.

Calculations based on Newtonian mechanics indicated that the slight perturbations that had been observed in the orbit of the planet Uranus could be accounted for by the presence of an additional planet in the solar system, one previously unknown.

In 1846, this new planet, Neptune, was found at the exact position in space indicated by the calculation.

By the middle of the 19th century, the ancient mystery of the nature of heat had been solved by the kinetic theory of matter and the science of thermodynamics.

The latter had also provided an understanding for the 'arrow of time' inherent in the workings of Nature.

Classical physics reached its zenith in the second half of the 19th century with the publication of Maxwell's theory of electromagnetism and the discovery of the electromagnetic waves whose existence it had predicted.

This theory summarized and unified everything that was known at the time about electrical and magnetic phenomena and provided the first comprehensive conceptual basis for the science of optics.

However, towards the end of the 19th century, it became clear that there were important physical phenomena for which classical physics had no satisfactory explanation.

The electron, X-rays and radioactivity, all of which were discovered within a few years of each other in the last decade of the century, were beyond the competence of classical physics to explain.

Moreover, there were even instances where the hypotheses and laws of classical physics were found to be totally incompatible with the results of experiments in more conventional fields, such as studies of the speed of light relative to different observers and the emission and absorption of heat radiation.

It became apparent that despite its great achievements, some of the most fundamental principles underlying classical physics were incorrect.

Clearly, fresh ideas and theories were needed.

At the beginning of the 20th century, three new hypotheses were put forward that changed the face of physics. They were:

1. **The theory of relativity;**
2. **Quantum theory;**
3. **The nuclear model of the atom.**

This triad laid the foundations of modern physics.

They and the effects derived from them are the subject of this class.

The new theories of modern physics not only resolved the problems left unanswered by classical physics but extended the reach of the physical sciences into previously unknown fields.

In general, the familiar and well established laws of classical physics remain valid for dealing with phenomena that occur on a ‘normal’ scale, in the human sense of the term.

However, when dealing with phenomena occurring on the cosmic scale on the one hand, and on the atomic scale on the other, only the more comprehensive laws of modern physics can be employed.

Our study of 20th century physics opens with the discovery of the electron, the first elementary particle to be identified.

This is followed by a short review of the classical theory of electromagnetic waves.

These two topics provide the basis for the comprehensive examination of the theory of special relativity that follows.

Part One of this class concludes with a brief survey of the theory of general relativity.

Part 1.1 The Electron

When first investigated in the 18th and 19th centuries, electrical and magnetic phenomena were generally construed in terms of ae*thereal fluids*, as were those associated with heat and light.

These fluids were thought to comprise minute mutually repelling particles.

Thus, heat was either thought to be vibrations in the fluid caloric or an accumulation of this fluid in the interstices of materials.

Light was either a flux of particles emitted at high *speed* from luminous bodies or the vibrations of a ubiquitous fluid aether.

Electric *fluids* - *electricity* - *flowed* readily through metals and other conductors but did not penetrate insulators such as paper and glass.

Opinion was divided as to whether there was just one electric fluid or two - a *positive* fluid and a *negative* fluid.

The possibility that electricity might not be a continuous fluid was first raised in the middle of the 19th century following Faraday's quantitative researches on electrolysis.

These showed the existence of a systematic relationship between the amount of electricity passed through an electrolytic cell and the quantity of material that undergoes chemical reaction (electrolysis) in the cell.

Thus, the passage of a certain amount of electricity - 96,500 coulombs in modern terms - always liberates a gram-equivalent of substance from the electrolyte, whether it is the metal released at the negative cathode or the non-metal substance at the positive anode.

Putting aside his misgivings about atomism, Faraday recognized that this suggested electricity might be atomic in nature and that a natural indivisible unit of electricity exists.

In 1891Stoney suggested that this natural unit of electricity be called an electron**.**

On this view, every ion carries an integer multiple of this natural unit.

For example, a silver ion, Ag^{+}, carries a single natural unit of positive charge; a typical copper ion, Cu^{+} $^{+}$**,**carries two such units.

Given that a gram-equivalent of monovalent ions comprises a mole, the magnitude of the natural unit of electricity, *e***,** can be calculated by dividing the 96,500 coulombs by Avogadro's number, 6.02 x10^{23}:

$$e = \frac{96{,}500}{6.02 \cdot 10^{23}} = 1.6 \cdot 10^{-19}\,\mathrm{C} \qquad (1.1)$$

The term *electron* is now used to designate the elementary particle that carries the natural unit of negative charge which was first identified towards the end of the 19th century in experiments on the conduction of electricity through gases at very low pressures.

At atmospheric pressure, gases do not usually conduct electricity.

However, at reduced pressures of O.5mm Hg to l0 mmHg and with applied potentials of several thousand volts, they can be made to pass a current.

These greatly reduced gas pressures were first achieved at the end of the 19th century following the advances made at that time in vacuum pump technology.

The gases were contained in narrow glass tubes, called discharge tubes, into which suitabIe electrodes had been inserted.

The passage of the current is accompanied by the appearance of striking colors in the tubes (Fig 1.1).

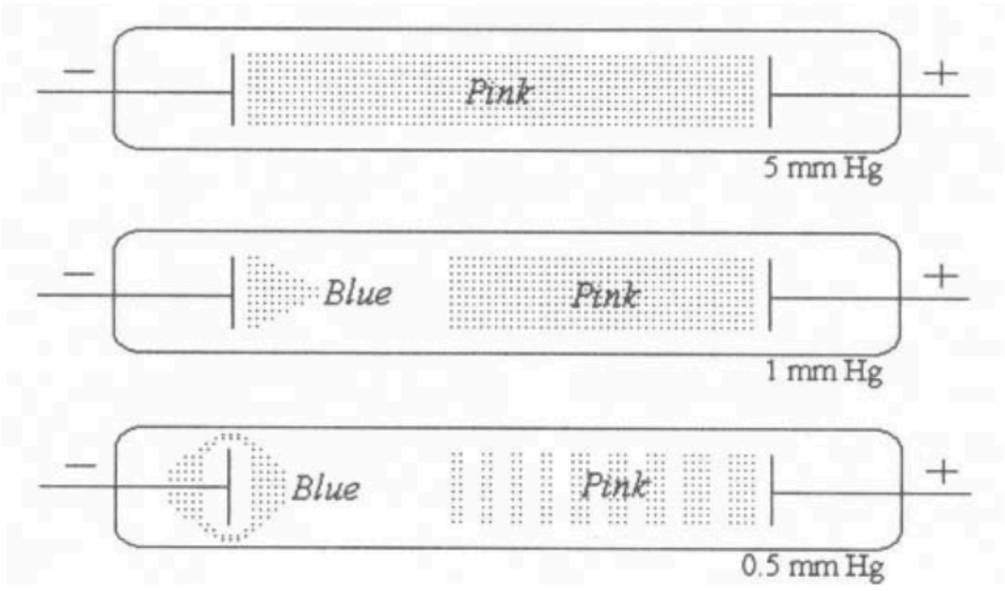


Fig 1.1 The conduction of electricity through air at low pressures in a discharge tube. The colors result from the excitation and ionization of the atoms of the gas in the discharge tube. The bands and the colored and dark regions arise from the variations in the electric field strength throughout the tube.

At still lower pressures, -0.02mm Hg, the colors disappear but the glass tube itself begins to glow with a green hue.

An object placed in front of the cathode (the negative electrode) casts a shadow on the opposite wall of the discharge tube (Fig 1.2).

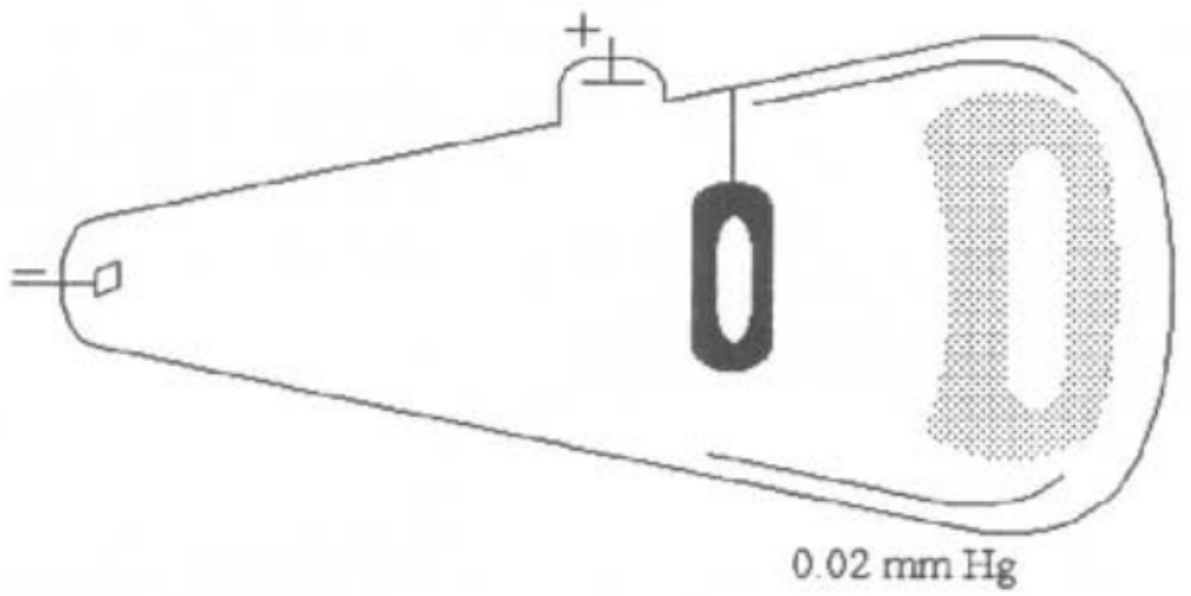


Fig 1.2 The shadow cast by an object in the path of the rays in a cathode ray tube.

Certain minerals, when placed in front of the cathode, fluoresce with brilliant colors. It appears that something is being emitted from the cathode; this emanation was given the name ***cathode rays.***

In further experiments it was found that the cathode rays were deflected by a magnetic field as would a stream of negative charge (Fig 1.3a).

Furthermore, a small paddle wheel positioned between the electrodes rotated under their impact; switching the polarity of the electrodes reversed the direction of the rotation (Fig 1.3b).

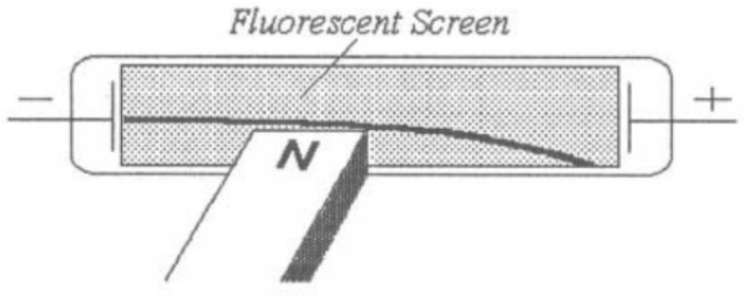


Fig 1.3a The deflection of cathode rays in a magnetic field. The direction of the deflection shows that it is negative charge that flows with the cathode rays.

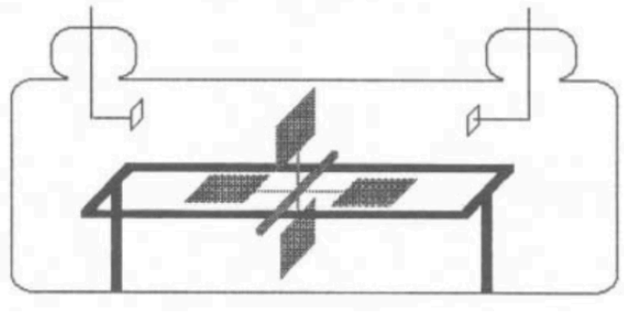


Fig 1.3b The paddle wheel rotates under the impact of the cathode rays.

These two phenomena suggested the cathode rays might be negatively charged particles.

Nevertheless, many physicists at the time still considered them to be of an Ethereal rather than a material nature.

Convinced that the cathode rays were in fact charged particles of matter, J.J.Thomson set out in 1897 to measure their velocity, v , and the ratio, their charge, q, and their mass, m.

In one of the experiments he conducted, a narrow collimated beam of cathode rays was aimed along the length of a very low pressure glass discharge tube.

After emerging from the hole in the anode, the beam passed through a thin slit, between a pair of vertical coils and, finally, between the horizontal parallel plates of a condenser.

The green spot that appeared on the glass at the far end of the tube indicated where the beam impinged upon it (Fig 1.4).

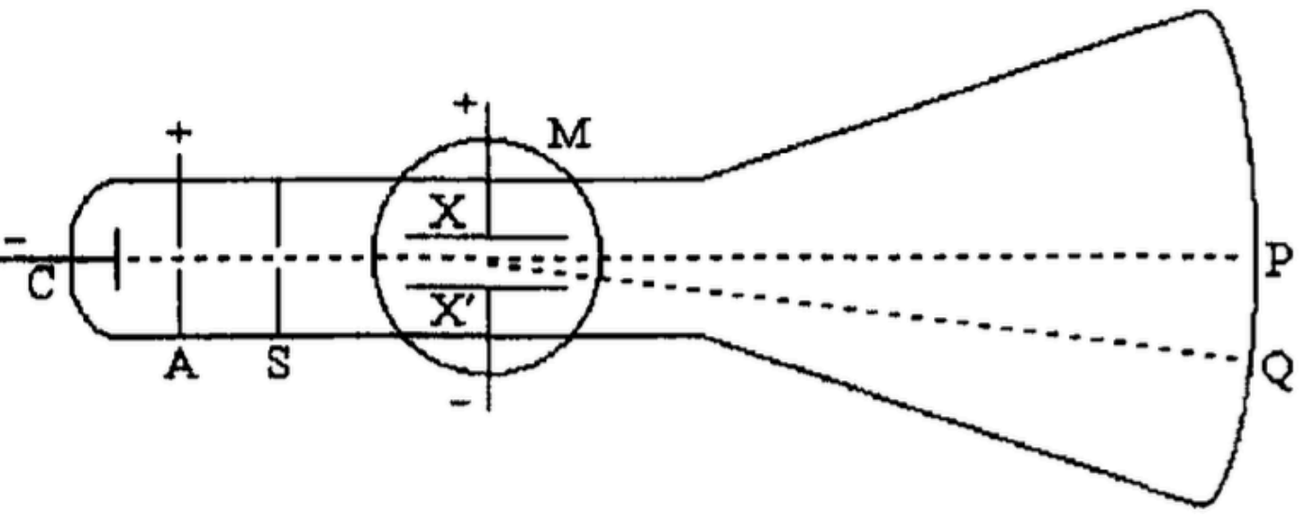


Fig 1.4 The type of very low pressure discharge tube used by Thomson for the determination of the ratio between the charge and the mass of the particles (electrons) in cathode rays. A collimated beam was produced by the arrangement of the cathode, C, the pierced anode, A, and the slit, S. The circle M represents the pair of coils, one on each side of the tube. Passing a current through these coils produces a uniform horizontal magnetic field in the gap between them. X and X' are the parallel plates of the Condenser. Connecting the plates to a source of a potential produces a uniform vertical electric field between them. An undeflected particle beam strikes the point P.

The cathode ray beam could be deflected vertically by two different means: (i) magnetically; (ii) electrically:

(i) Magnetically—by passing a current through the coils. This induces a uniform horizontal magnetic field, B, in the space between them which, in turn, exerts a force, F_B, on the cathode ray particles such that

$$F_B = Bqv \tag{1.2}$$

where v is the velocity of the particles. This force acts at right angles to the direction of the particle's motion and so constitutes a centripetal force. Thus, as they pass between the pair of coils, the cathode ray particles move in a circular path—along the arc of a circle—such that

$$F_B = Bqv = {}^{mv^2}\!/_{r} \tag{1.3}$$

where r is the radius of the arc.

(ii) Electrically—by connecting the condenser plates to a potential source; this produces a vertical electric field, E, in the space between them that exerts a vertical force, F_E, on the cathode ray particles such that

$$F_E = Eq \tag{1.4}$$

Initially, Thomson applied just the magnetic field.

This had the effect of moving the green spot down from P to Q.

He then activated the electric field and adjusted the potential between the condenser plates until the field was just sufficient to return the green spot from Q to P.

At this point, the two forces, F_B and F_E, cancelled each other out, such that:

$$F_E = F_B$$
$$\text{or} \quad qE = Bqv = {}^{mv^2}/_{r} \tag{1.5}$$

From which, the particles' charge to mass ratio is given by

$${}^{q}/_{m} = {}^{E}/_{rB^2} \tag{1.6}$$

Thomson knew the value of the magnetic field, B, from the dimensions of the coils and the current flowing through them, of the arc radius, *r,* from the dimensions of the apparatus and the distance PQ and of the electric field, E, from the potential applied between the condenser plates and the distance between them.

Substituting these in equation (1.6) gave a value, in modem SI units, of -2×10^{11}C/kg for the charge to mass ratio of the cathode ray particles.

Thompson repeated the experiment using different metal cathodes and with different gases in the low pressure tube, but in each case he found approximately the same value for the ratio.

The value found by Thomson for the charge to mass ratio of the cathode ray particles was three orders of magnitude greater than the largest previously known value of this ratio, namely, that found for the aqueous hydrogen ion, $H^{+}_{(aq)}$.

This could be attributed either "to the smallness of *m* or the largeness of *q* , or a combination of these two".

Thomson opted for the "smallness of *m*" and assumed that the magnitude of the charge, *q*, carried by the cathode ray particles was equal to the smallest charge known to be carried by any ion, i.e., $q = -1.6 \times 10^{-19} C$.

On this assumption, he calculated the mass of the cathode ray particles and obtained the result, astonishing at the time, that their mass was only 1/1836 that of the hydrogen atom, itself the smallest of all atoms.

In an attempt to verify Thomson's assumption, direct measurements of the magnitude of the charges carried by gaseous ions were made.

Although these experiments suffered initially from many sources of error, their results appeared to confirm Thomson's supposition.

The issue was finally and unequivocally settled in 1906 when, in a series of accurate and careful experiments, Millikan measured the magnitude of the electrical charges carried by both positively and negatively charged oil droplets.

In none of the hundreds of measurements he made, did he detect a droplet that carried a charge whose magnitude was less than 1.6×10^{-19}C , nor one that carried a charge whose magnitude was a fractional multiple of this amount.

In every case, the magnitude of the charges was an integer multiple of $e = 1.6 \times 10^{-19}$C .

The clear inference was that electrical charge-positive or negative-always appears as an integer multiple of the natural unit *e*.

The cathode ray particles Thomson had discovered are the elementary particles we now call electrons; the electron was the first elementary particle to be discovered.

Electrons carry the natural unit, -e, of negative charge.

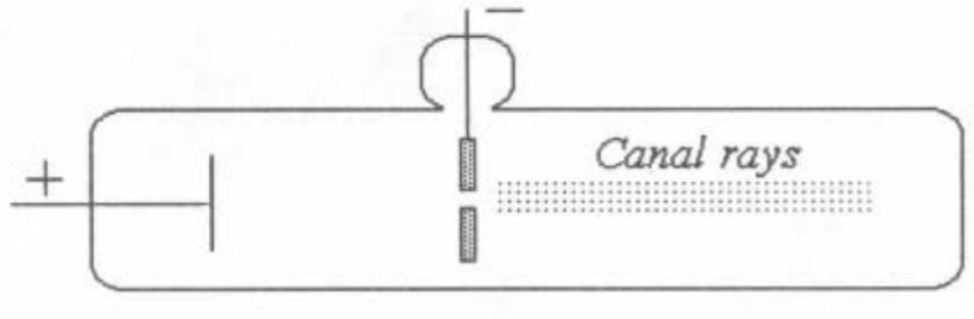
They are the smallest of the three constituent particles--protons, neutrons and electrons--of ordinary matter.

Though electrons have no discernible internal structure or dimensions, they nevertheless possess an intrinsic angular momentum (spin) and an associated magnetic moment.

Cathode rays were not the only emanations observed in discharge tubes.

If the cathode in the discharge tube was pierced, rays were seen to emerge from the hole in the direction away from the anode (the positive electrode), i.e., in the opposite direction to the cathode rays (Fig 1.5).

Fig 1.5 Positively charged canal rays emitted from a hole in the cathode.



These emanations were called canal rays; they were found to be positively charged.

Having identified the cathode rays, Thomson proceeded to investigate the nature of the canal rays.

These too proved to be particles of matter.

By measuring the ratio between their charge and their mass he identified them as positive ions of the gas present in the discharge tube.

He construed that these positive ions were produced in the space between the cathode and the anode from the bombardment of the gaseous atoms by the cathode rays (electrons).

Since they carried a positive charge, they were attracted to and accelerated towards the negative cathode, passing through the hole that had been pierced in it and emerging from the other side.

In one notable experiment, in which the gas in the discharge tube was a pure sample of the noble gas neon, Thomson discovered two different species of positive ions in the canal rays whose masses differed by about 10%.

The ions could only be neon ions, the purity of the sample ensured this.

This discovery of different species of neon ions first demonstrated the existence of atoms that share the same chemical identity but have different masses; what we now refer to as isotopes.

In 1932, a particle was discovered that is identical to the electron in all its properties except that it is positively and not negatively charged; it was designated the anti-electron or positron.

When a positron and electron collide, the pair of particles self-destruct, releasing a burst of high frequency electromagnetic radiation - γ rays.

Part 1.2 Electromagnetic Waves

The notion that an aether pervades all of space, including even that occupied by material bodies, was popularized in the 17th century by the philosopher Descartes.

Traditionally, space had been regarded as passive and unchangeable; an inert background, like a cinema screen, upon which any action can be projected but without which no action can be perceived.

Descartes argued that in order to account for many physical phenomena, a dynamic active medium - *an Ether*- that transmitted the forces exerted at a distance by one object on another, had to be postulated.

It was also generally accepted that the propagation of light involved an Ether.

However, two conflicting hypotheses held sway at various times concerning the nature of light and the role of this Ether in its propagation.

One theory asserted that light is a particle or corpuscular phenomenon: a flux of aethereal particles emitted at high speed from luminous bodies.

The other, that light is a wave phenomenon: the vibrations of a ubiquitous luminiferous ether.

According to the corpuscular theory - the one advocated by Newton - light propagates at a greater speed through a dense medium than through a less dense one.

On this view, the speed of light in water should be greater than in air.

The wave theory proposed by his 17th century contemporary Huygens posited just the opposite: that light should propagate faster in air than in water.

Fairly good estimates of the speed of light-about 200,000 km/s-had been obtained at the time by Romer from astronomical observations.

However, the technical means required for a direct terrestrial comparison of the speed of light in air and water and for carrying out an experimentum *crucis* that would decide between the two theories would not be available for almost 200 years.

Until then, the dispute between the two hypotheses was conducted as much on the basis of the eminence of their respective supporters as on their relative capacities to explain the phenomena.

The strength of Newton's corpuscular theory lay in the simple explanation it provided for the rectilinear propagation of light, i.e., for the formation of well-defined shadows.

It postulated that the material corpuscles that constitute a beam of light are emitted from luminous bodies in straight lines, as required by Newton's first law of motion, and hence the shadows; that they are reflected elastically from mirrors, which accounts for the familiar laws of reflection; and that they vary in size according to the color of the light.

In rejecting this approach, Huygens pointed to the extreme speed at which Iight propagates and to the fact that when light rays intersect they cross without hindrance and without leaving any lasting impression on one another.

Material particles could not possibly travel at such a speed; nor could their paths cross without the corpuscles colliding and deflecting one another.

Hence, he argued, it must be by some other mechanism that light propagates.

In his opinion, the propagation of light was to be understood by analogy to the propagation of sound.

Throughout the 18th century, Newton's corpuscular theorxy of light was the one embraced by most scientists.

However, by the middle of the 19th century, opinion had completely changed and the issue appeared to have been settled once and for all in favor of the wave hypothesis.

The turning-point was the landmark experiment conducted by Thomas Young in 1801.

Young observed that when light from a single source is projected onto a white screen through two adjacent pin-holes or narrow slits, a sequence of bright and dark bands-fringes-appears on the screen (Fig 1.6).

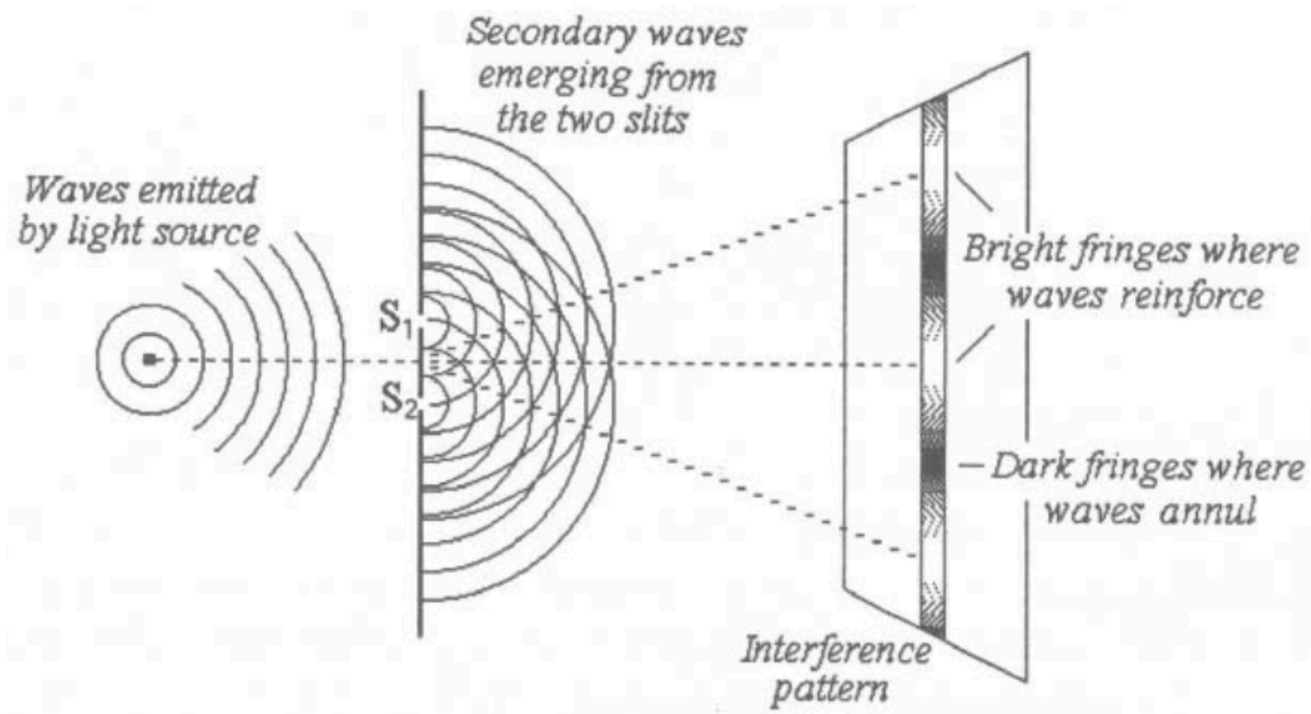
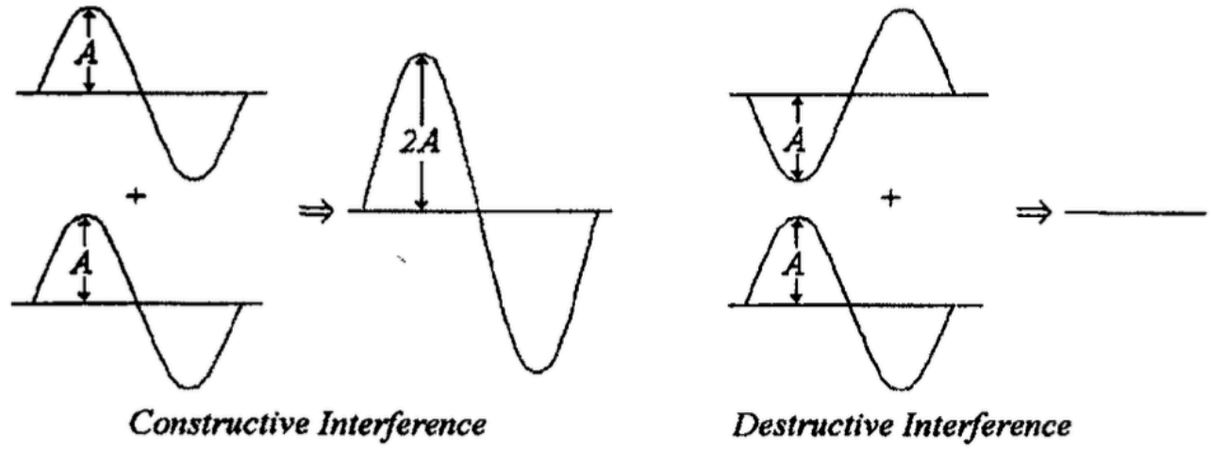


Fig 1.6 Young's double-slit experiment. Light from the source passes through the narrow slits, S_1, and S_2, and emerges as two coherent trains of secondary waves. Young explained that bright fringes appear on the screen where the secondary waves reinforce each other (the wave-trains are in phase with each other) and dark fringes where they annul one another (the wave-trains are out of phase).

In order to account for the appearance of the fringes, Young proposed the simple hypothesis that if two waves are superposed, their joint effect is the combination of the two.

In so doing, he provided the key concept-interference or superposition-needed for the establishment of the wave theory of light (Fig 1.7).

Fig **1.7** The destructive and constructive interference of two waves. If light waves are in phase when or where they join, they reinforce each other and produce light of greater intensity (constructive interference); if they are out of phase, they annul one another and produce darkness (destructive interference).



He wrote:

Suppose a number of equal water waves move upon the surface of a stagnant lake at a constant velocity and enter a narrow channel leading out of the lake. Suppose another similar series of waves arrive at the same channel with the same velocity. Neither series of waves will destroy the other; but their effects will be combined: if they enter the channel such that the elevations of the one coincide with those of the other; they must together produce a series of greater joint elevations [constructive interference]; but if the elevations of one series correspond to the depressions of the other, they must exactly fill up these depressions and the surface of the water must remain smooth [destructive interference] . Now, I maintain that similar effects take place whenever two portions of light are thus mixed.

Following Young's experiment and the subsequent mathematical elucidations of the interference and diffraction of light based on his principle of superposition, it was generally accepted that "light is in the ether, the same thing as sound in the air", i.e., light is a mechanical wave in the ether, just as sound is a mechanical wave in the air.

However, further efforts to clarify the actual physical nature of light waves were confounded by the phenomenon of light polarization.

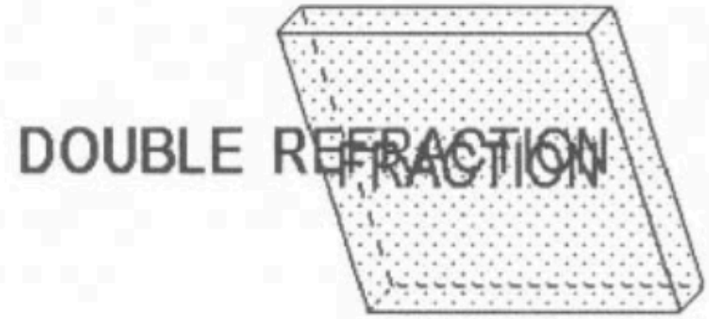


Fig **1.8a** Doublerefraction. When looking at a printed page through a calcite crystal, two images are seen.

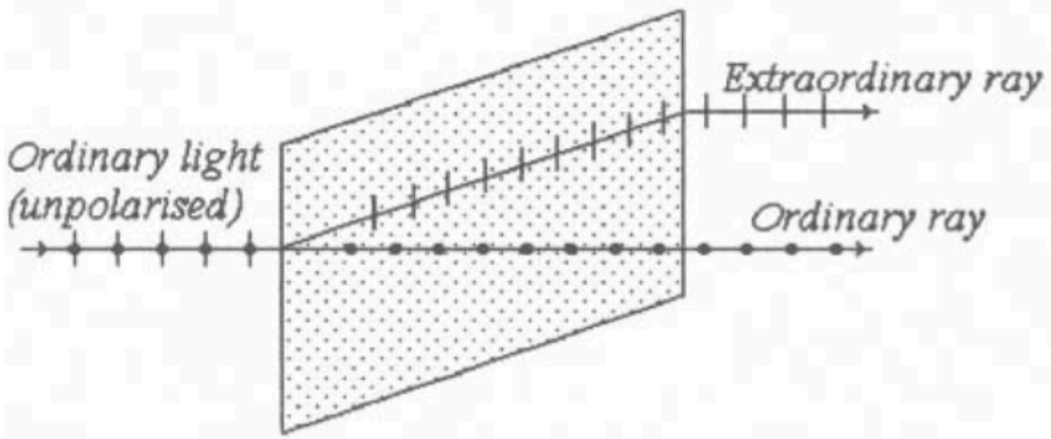


Fig **1.8b** An incident ray of ordinary (unpolarized) light is split into two rays on passing through a calcite crystal. The ordinary and extraordinary rays are polarized at right angles.

It had been known for some time that when a beam of ordinary light passes through a calcite crystal it splits into two beams; this is the phenomenon known as double refraction or birefringence.

If the two beams are then passed through a second calcite crystal they do not behave like ordinary light.

Furthermore, the two beams are in turn extinguished when viewed through a rotated tourmaline crystal (or polaroid).

We say that the two light-beams are polarized at right angles (Fig 1.8).

In 1808, Malus discovered that such polarized light can also be obtained by the reflection of ordinary light from a glass surface.

When differently polarized beams of light are joined, no interference pattern is produced.

In order to account for this and for the other phenomena associated with light polarization, Young and Fresnel determined that light waves must be transverse ether waves, i.e., the ether particles vibrate at right angles to the direction of light propagation.

Sound waves in air are longitudinal waves, i.e., the air particles vibrate parallel to the direction of the sound propagation (Fig 1.9).

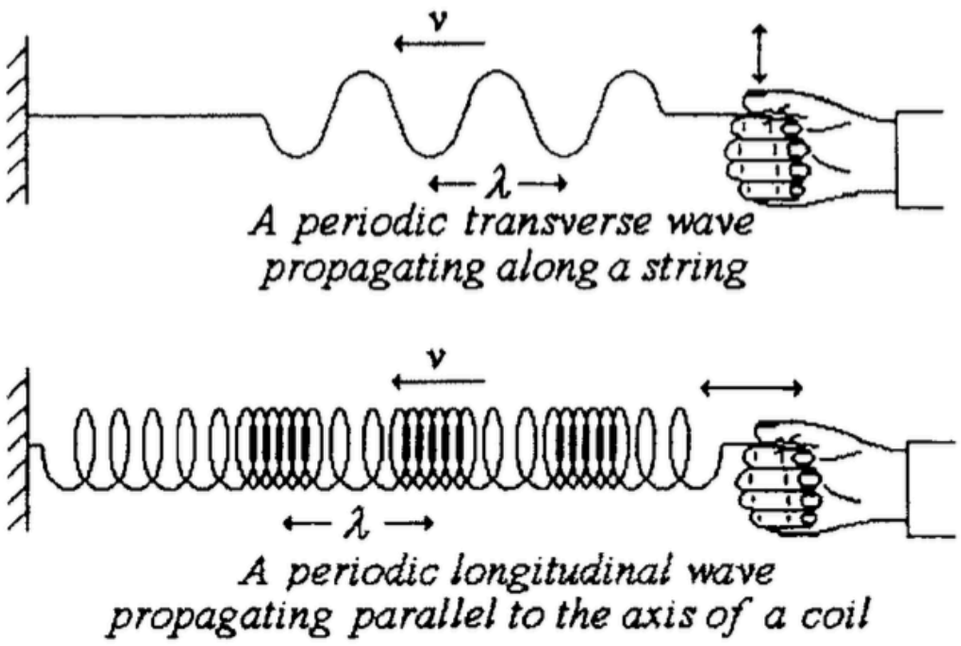


Fig 1.9 Transverse and longitudinal periodic waves. Moving the hand up and down produces vertical transverse waves along the string; moving it sideways produces horizontal transverse waves. Indeed, transverse waves can be generated in any plane. Such transverse waves are plane polarized. However, if the coil is pulled to and fro, the longitudinal waves produced can only propagate in one way, parallel to the axis of the coil. In both cases, the frequency of the waves, f, is set by that at which the hand oscillates and their wavelength, λ, by the ratio between the velocity at which they propagate, v, and their frequency: $\lambda = v/f$

Polarization is not exhibited by longitudinal waves.

Thus, light polarization has no sound analogue.

Transverse mechanical waves cannot propagate through fluids such as water or air but only through solids.

Consequently, if the light waves that propagate through space are indeed transverse aether waves, the aether must be described in terms of an elastic solid.

Transverse waves propagate through elastic solids with velocity

$$v = \sqrt{S/\rho} \tag{1.7}$$

where S is the shear modulus of the solid and ρ is its density.

The speed of light was known to be of the order 10^8m/s.

For the velocity of the ether waves to be of this magnitude, the ether had to be more rigid than steel (its shear modulus S had to be very large) whilst at the same time its density had to be less than that of air (ρ very small)!

Although the existence of a material with such contradictory properties was quite inconceivable, the concept of a luminiferous ether was still not abandoned.

The aether concept was further complicated when Faraday, having conceived the notion of electric and magnetic fields, successfully employed the idea of lines of force in explaining electrical and magnetic interactions and in formulating the laws of electromagnetic induction.

Faraday's ideas were not universally accepted.

There were those who argued that the forces acting at a distance between charges do so directly, being the result of an instantaneous unmediated interaction, the same way as it was thought the force of gravity acted.

notion can be represented schematically as:

Charge ↔ Charge

However, in his revolutionary elucidation of electrical and magnetic interactions, Faraday placed the emphasis on the space between the interacting bodies.

He asserted that the space surrounding a magnet is “a part of the true and complete magnetic system”.

In Faraday’s view, space plays an active role in the propagation and transmission of the interactions.

As Maxwell wrote:

> *Faraday, in his mind’s eye,* saw *lines of force traversing all space where the mathematicians* saw *centers of force attracting at a distance; Faraday saw a medium where they saw nothing but distance; Faraday sought the seat of the phenomena in real actions going on in the medium*.

According to this approach, electrical and magnetic forces act through the medium of a field that pervades the space between the charges or the magnets.

In order to sustain this field and its lines of force, it was assumed that something must be present in the space, Faute de mieux, Faraday presumed this to be the ether.

In support of his view, Faraday cited the fact that though polarized light is normally unaffected by its passage through ordinary glass, when a beam of polarized light passes through a glass-block in a magnetic field, the form of its polarization changes.

Presuming light to be an aether wave, the conclusion Faraday drew from this was that the lines of force of electric and magnetic fields could be regarded as elastic deformations in the aether.

Thus, both electrical and magnetic properties were now added to the exotic mechanical properties already required of the ether.

Convinced by his reading of Faraday's ***Experimental Researches in Electricity*** that "his method of conceiving the phenomena was also a mathematical one, though not exhibited in the conventional form of mathematical symbols", Maxwell took upon himself the task of providing a firm mathematical basis to Faraday's intuitive ideas.

In an immense enterprise, he developed a mathematical system that described and accounted for all the then known electrical and magnetic phenomena in terms of electric and magnetic fields.

In doing so, Maxwell formulated two fundamental postulates:

1. **that a moving or changing magnetic field sets up an electric field;**
2. **that a moving or changing electric field sets up a magnetic field.**

The mathematical functions and equations Maxwell developed from these two symmetrical assumptions predicted the existence of transverse electromagnetic waves, i.e., disturbances in the electric and magnetic fields that would propagate through empty space with a velocity, ***c*** **,** given by the formula:

$$c = \sqrt{1/\mu_0 \varepsilon_0} \tag{1.8}$$

where μ_0 is the permeability of empty (free) space and ε_0 is its permittivity.

The numerical values of these fundamental constants were already known from electromagnetic experiments.

Hence, calculating the velocity of propagation of the electromagnetic waves just required their substitution in equation (1.8).

Maxwell wrote:

> *If it should be found that the velocity of propagation of electromagnetic disturbances is the same as the velocity of light ... we shall have strong reasons for believing that light is an electromagnetic phenomenon.*

Substituting the numerical values for the permeability and permittivity of empty space, viz. $\mu_0 = 4\text{x}10^{-7}$ m kg C^{-2} and $\varepsilon_0 = 8.85\text{x}10\text{-}12$ C^2 $N^{-1}m^{-2}$, in equation (1.8) does indeed give a wave velocity equal to the measured speed of light in vacuo, c $=3\text{x}10^8$ m/s.

$$\sqrt{\frac{1}{\mu_0 \varepsilon_0}} = \sqrt{\frac{1}{\left(4\pi \cdot 10^{-7}\right) \cdot \left(8.85 \cdot 10^{-12}\right)}} = 3 \cdot 10^8 \text{ m/s}$$

Maxwell inferred from this identity that light itself is an electromagnetic wave.

Maxwell summarized his ideas in ***A Treatise on Electricity and Magnetism*** that was published in 1873.

Fifteen years later, Hertz succeeded in transmitting and receiving the periodic electromagnetic waves that, according to Maxwell's theory, should be produced by oscillating electrical circuits.

His experiments confirmed all of Maxwell's predictions.

Maxwell's equations summarize all the laws of electromagnetism just as Newton's laws do for classical mechanics.

It should be noted that these equations do not involve any actual assumptions about the existence of the aether.

However, in his total rejection of unmediated action at a distance, Maxwell continued to believe in its existence.

He regarded it as being the medium that carried the potential and kinetic energy associated with electromagnetic waves and even visualized electromagnetism as a kind of aether mechanics, imagining space to be filled with 'aethereal' interlocking gear wheels that transmitted the electromagnetic forces.

1.2.1 The Production and Properties of Electromagnetic Waves

In general, all accelerated electric charges radiate electromagnetic energy in the form of electromagnetic waves.

If the charge constitutes an electric oscillator, the energy is radiated as periodic electromagnetic waves whose frequency is equal to that of the oscillating charge.

Similarly, electric oscillators possessing a natural frequency that matches that of the electromagnetic radiation incident upon them, i.e., when there is resonance between the oscillators and the incident radiation, will absorb energy from the radiation.

The electric and magnetic disturbances that constitute an electromagnetic wave propagate in phase with one another but at right angles to each other (Fig 1.10**).**

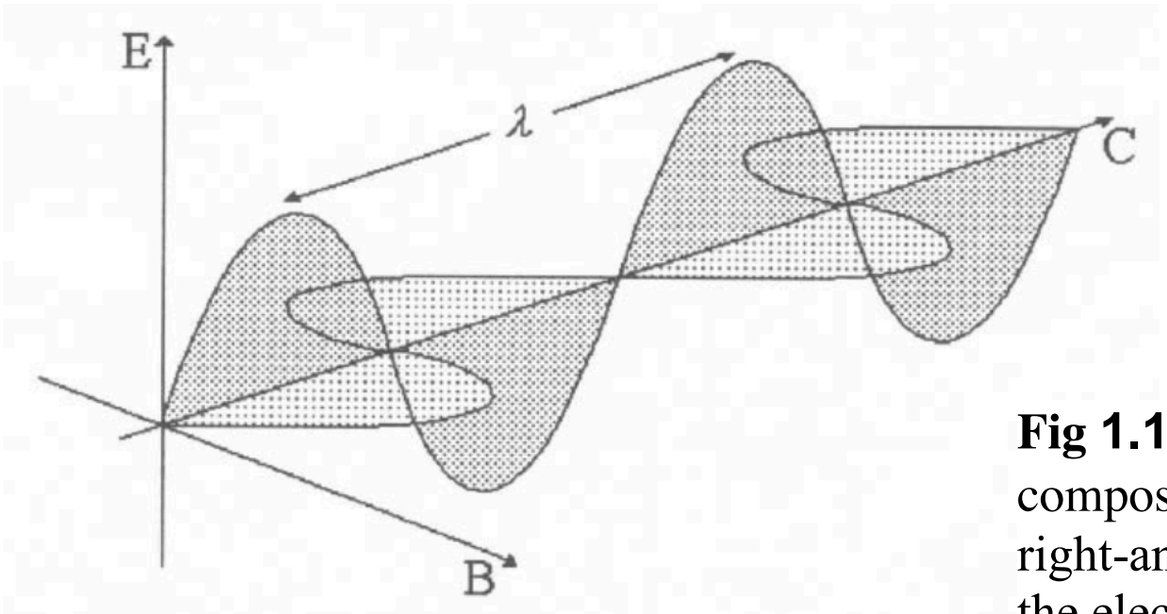


Fig 1.10 A 'snapshot' of a periodic electromagnetic wave. The wave is composed of a periodic electric field and a periodic magnetic field at right-angles to one another but in phase with one another. E indicates the electric field, B the magnetic field and C the direction of propagation of the wave.

The characteristic property of a periodic electromagnetic wave is its frequency; this is the parameter that fixes its color**.**

A continuous spectrum of periodic electromagnetic waves has been identified with frequencies ranging from about 10^4Hz to 10^{22}Hz(Fig 1.11).

The waves comprising the electromagnetic spectrum all propagate through empty space at the speed of light, c = $3x10^8$ m/s.

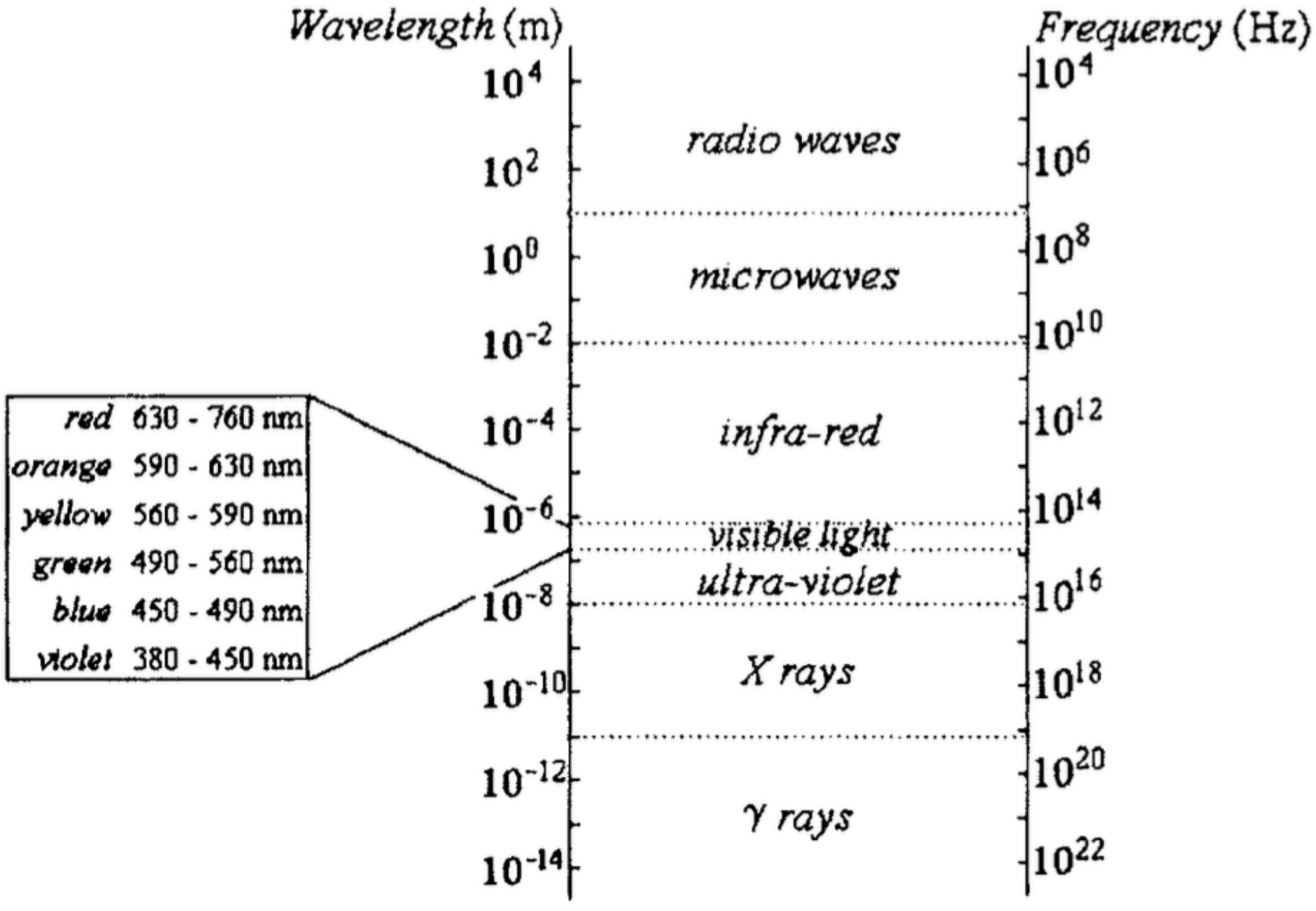


Fig 1.11 The electromagnetic spectrum. The values given for the wavelengths are those exhibited by the radiation when it propagates in vucuo.

The lower frequencies of the electromagnetic spectrum, such as radio waves, are generated by oscillating electrical circuits; these frequencies are in turn detected (absorbed) by the use of aerials in which the incident radiation sets up corresponding electrical oscillations.

Microwaves, such as those used for radar detection and rapid cooking (microwave ovens), are generated by the velocity modulation (acceleration) of electron beams.

The emission and absorption of infra-red radiation (heat radiation), visible light and ultra-violet radiation usually involves the transfer of electrons between the energy levels associated with the outer structure of atoms, molecules or ions.

X rays are generated by the deceleration of high energy electrons or by electron transfers between the energy levels associated with the inner structure of atoms.

The high frequency γ rays are emitted during the decay of radioactive nuclei.

1.2.2 The Limits of Electromagnetic Theory

Despite its outstanding success in elucidating electrical, magnetic and optical phenomena, it was subsequently found that there were certain physical phenomena for which Maxwell's theory could not account.

In particular, it failed to explain many of the phenomena involving interactions between electromagnetic radiation and matter.

The discovery of the electron as the 19th century drew to its close was also totally unexpected.

When considering what happens in a wire carrying an electric current Faraday had written:

> *By a current I mean anything progressive, whether it be a fluid of electricity or two fluids moving in opposite directions or merely vibrations, or speaking more generally, progressive force.*

This nebulous view of an electric current was at the root of Maxwell's postulate that an electric current (a displacement current) flows through the space between the plates of a condenser as it is charged and gives rise to magnetic effects equivalent to those produced by an ordinary current flowing through a wire.

The shock of the discovery of the electron to the followers of Faraday and Maxwell is well illustrated by the following contemporary remark:

> *The separate existence of a detached atom of electricity never occurred to me as possible. And if it had, and I had openly expressed such heterodox opinions, I should hardly have been considered a serious physicist. For the limits to allowable heterodoxy in science are soon reached.*

There remains to this day a troubling and still unresolved aspect of Maxwell's equations, viz. their indifference to the distinction between past and future.

The electromagnetic waves whose existence the equations posit can travel backwards or forwards in time; the equations do not tell us whether waves arrive before or after they were emitted.

Because we never seem to observe radio waves traveling backwards in time, the backward-in-time solutions of the equations are usually dismissed as being ‘unphysical’.

Nature exhibits an intrinsic awareness of past-future, the so called *Arrow of Time*, that has yet to be satisfactorily incorporated into scientific theories.

Part 1.3 The Special Theory of Relativity

The aethcr enigma was first resolved in 1905 with the publication of Einstein’s special theory of relativity.

This revolutionary hypothesis redefined some of the most basic assumptions of the physical sciences.

To fully appreciate the nature of these new definitions and of the circumstances that made them necessary, we must return to the very fundamentals of science and reconsider the rationale that underlies our individual and collective awareness of the physical world around us.

Does a physical world, independent of humanity, really exist?

Do mountains, trees, land, sea and sky all actualIy exist whether or not anybody is looking at them?

After all, our awareness of the external world occurs within us, in our minds; perhaps it is all an illusion.

George Berkeley's (1708) version of solipsism denies the existence of anything outside the mind

> *When we do our utmost to conceive the existence of external bodies, we are all the while contemplating our own ideas. But the mind, taking no notice of itself is deluded into thinking it can and does conceive bodies existing with aught of or without the mind.*

Furthermore, since no experiment can prove the physical existence of an object that is indiscernible to human observers, what meaning can *physical existence* have in the absence of such observers?

On the other hand, does everything we observe or perceive really exist?

Can we always distinguish between real and virtual reality?

Where is the border between reality and illusion?

The prototypes of the resolutions Western Civilization has contrived for this mystery are the two hypotheses associated with the classical Greek philosophers, Plato and Aristotle, depicted in Raphael's' Vatican fresco *The School of Athens*.

The one linked with Plato's name asserts that all true knowledge is to be found through the mind's contemplation of abstract ideas.

> *Plato admitted the existence of an external world but came to the conclusion that the world perceived by the senses is motley, protean, ever-changing and unreliable. The true world is the world of ideas, which is unchangeable and incorruptible. Yet this world* ***of*** *ideas is not accessible to the senses but only to the mind. Observations are useless.*

In Plato's view, physical evidence proves nothing; it cannot overturn pure reasoning.

The reality behind appearances, what is inherently true of them, is the pure reason of mathematics.

Man interprets and remoulds the imperfect world he experiences in terms of his vision of perfection.

The idea of lines, circles and squares is more real than physical nature, in which there are few pure examples of these perfect forms.

The alternative view, that associated with Aristotle, holds that it is material things that are the primary substance and source of reality.

> *Physics, and science generally, must study the physical world and obtain truths* from *it. Thus, genuine knowledge is obtained from sense experience by intuition and abstraction. These abstractions have no existence independent of the mind. To arrive at truths, Aristotle used what he called universals - general qualities that are abstracted from real things.*

Using intuition and the human faculty for abstraction, Aristotle elevated the observed qualities of objects into mental concepts.

For example, since every material thing he saw around him undergoes corruption, decay or mortality, Aristotle concluded that *ceaseless change* must be a universal attribute of the sublunar world.

After dominating Western thinking for almost two millennia, the inadequacy of these ancient hypotheses became apparent at the Renaissance.

In their classical conceptions, neither the Ideas of Plato nor the Universals of Aristotle could cope with the new knowledge and the profusion of discoveries, inventions and innovations that, from the fifteenth century on, swept through Western Europe.

The new arts and technologies were simply beyond the scope and grasp of these hallowed doctrines.

Perceiving no limit or end to the flood of new ideas, the 17th century philosopher Descartes was forced to acknowledge that all the beliefs he presently held might perhaps be false.

He found just one fact of which he could still be sure - *Cogito ergo sum* - I think, therefore I am.

This truth made him aware of his own limitations.

But, by the same token, it also allowed him to conceive the existence of an Omnipotent Perfection.

Descartes argued, that this Perfection would not be a deceiver; He would not make us believe there was a material universe if it did not really exist.

Descartes averred that the material universe really does exist.

Moreover, and most important of all, he asserted that this universe works like a great machine, according to laws accessible to human thought and understanding.

Assured of its consonance and reality, the Europeans now set out in earnest to learn all they could about the physical world.

Their findings appeared to confirm Descartes' beliefs: the physical world was evidently both fathomable and governed by natural laws.

Thus, tradition and scholarship were replaced by reason and experiment as the arbiters of truth.

As Newton wrote in the opening passage of his ***Opticks,*** the first science text to be written in a language that could be understood by non-specialists:

> *My design in this book is not to explain the properties of light by hypotheses, but to propose and prove them by reason and experiments,*

If the phenomenon of light, the fiat of the very first day of Creation, could be explained by reason and experiment, was there anything in the world that could not?

The agenda of modern times had been set.

Measurement became the yardstick of reality.

A concept was considered physically real if something corresponding to it in the physical world could be measured.

Heat, though unseen, is physically real because we can measure temperature with a thermometer.

Similarly, color acquired a physical reality when Newton declared that “Lights which differ in Color, differ also in Degrees of Refrangibility - their disposition to be refracted or turned out of their Way in passing out of one transparent Body or Medium into another”.

In short shrift, every concept that did not satisfy the measurement criterion was purged from the body of science.

Only those for which a method or instrument of measurement could be devised were allowed to remain.

1.3.1 The Principle of Covariance

The foundation-stone upon which the new sciences were established was the belief that the physical world around us is real and that its existence is independent of any percipient entity.

On this view, physical changes are events, not the actions of an Agency; the physical world is driven by detectable intrinsic forces, inherent in its nature, not by unverifiable external causes; and all true knowledge ***is*** near at hand, reproducible and just waiting to be learnt.

Before uncritically accepting this maxim, you might like to ponder the fact that in modern information theory, which is concerned with that area where physical science and psychology overlap, the mathematical function that defines information is related to the thermodynamic function entropy in very fundamental ways.

Although information is a measure related to the observer and entropy is related to the physical system, the distinction between them is rather artificial since thermodynamics always implies an observer who establishes or removes constraints on the system or otherwise manipulates it.

> *Information deals with the usefulness of a set of symbols to an observer* ... *What appears to lie behind all this is the fact that our view of the physical universe* is *not a cold, lifeless abstraction but is structured* by *the human mind and the nature of life itself. While* it *has long been fashionable to inquire how biology depends on physics,* we *are being forced to inquire how physics depends on the underlying biology of the* mind. - Harold J. Morowitz,*Entropy for Biologists*(1970).

A corollary of this new vision of reality was that all observers of the physical world are, in principle, equally authoritative and legitimate.

All have the same access to knowledge.

No observer, whoever or wherever he is, has a better view of physical reality than any other; all observers are equivalent.

For example, each player in a football team constantly monitors the movements of the other players and of the ball.

But is any one player’s description of these motions superior to those of the others?

Obviously not.

Similarly, the description of planetary motions from the point of view of a stationary sun given by the Copernican heliocentric system, is not inherently superior to that of the Ptolemaic geocentric system which view things from a stationary Earth.

Nor is the opposite true: the Ptolemaic description is no better than the Copernican.

In fact, any motion described by the one can be replicated and accounted for by the other.

Given sufficiently complex mathematical devices, the celestial motions can be accurately and precisely described from anywhere.

As a consequence of their underlying equivalence, all observers should discover that the physical world is governed by the same laws; the laws of nature must take the same form for all observers.

Physical laws which take the same form for all observers, irrespective of their position or motion, are said to be ***covariant.***

A quantity whose measured value is unaffected by the position or motion of the observer is said to be ***invariant.***

Thus, the laws of nature must be covariant.

Furthermore, if all observers are subject to the same laws of nature, then these laws cannot be the basis for distinguishing between them; observers who are subject to the same laws are indistinguishable by reference to those laws, i.e., *individualism*, the sense of self, is essentially non-scientific.

Notwithstanding its apparent simplicity and self-evidence, the rigorous application of observer equivalence - the ***principle of covariance -*** to the formulation of the laws of physics has proved one of the greatest intellectual challenges of all times.

As we shall see**,** it was in order to correct the failure of classical physics to meet this challenge that Einstein put forward his theories of relativity.

A limited form of covariance is found in the following corollary Newton added to his laws of motion:

> *The motions of bodies included in a given space are the same among themselves[relative to one another], whether that space is at rest, or moves uniformly forwards in a right [straight] line… A clear proof of this we have from the experiment of a ship; where all motions happen after the same manner whether the ship is at rest, or is carried uniformly forwards in a right [straight] line*

Relative to an observer on board the ship, the motion of an object is the same whether the ship is stationary or sailing at a uniform velocity in a straight line.

In both cases, an object thrown vertically upwards falls straight back into the observer's hands; the motion of the object is subject to the same laws.

Thus, by reference to the motions they observe in the immediate space around them, observers at rest cannot be distinguished from those in uniform motion in a straight line; the latter can never be sure whether it is they or their surroundings that is moving.

Such indistinguishable observers are called *inertial observers*.

The notion of indistinguishable inertial observers for whom neither absolute rest nor absolute uniform motion has any meaning is the underlying postulate of Newtonian mechanics.

Newton identified an object's ***mass*** with the "quantity of matter" that lies in it: its inertia *(vis inertia)* or tendency to persist in a state of rest or uniform motion in a straight line.

This innate "force of matter" is an absolute; an invariant quantity, that is manifested to the same degree whenever an external force is impressed upon the object.

Thus, all inertial observers must exert the same force to effect a given change in an object's motion.

It follows that force too is an invariant quantity.

Newton's laws of motion are true for all inertial observers.

However, in their canonical form, they are not valid for non-inertial observers such as passengers in an accelerating or breaking vehicle or persons riding on a rotating carousel.

For example, according to Newton's First Law of Motion:

> *Every body continues in its state of rest, or of uniform motion in a right [straight] line, unless it is compelled to change that state by forces impressed upon it*.

i.e., a body's motion remains unchanged unless an external force acts upon it.

Observers on board a stationary ship or one that is sailing in a straight line at constant velocity find this to be true; persons and objects placed on deck stay where they are (so long as the ship remains on an even keel).

However, a vehicle's passengers are thrown, respectively, backwards or forwards every time it accelerates or brakes, even though no actual external force pushes them.

Similarly, relative to an observer on a rotating carousel, a free object moves in a curved path even though no actual external force acts upon it (Fig 1.12).

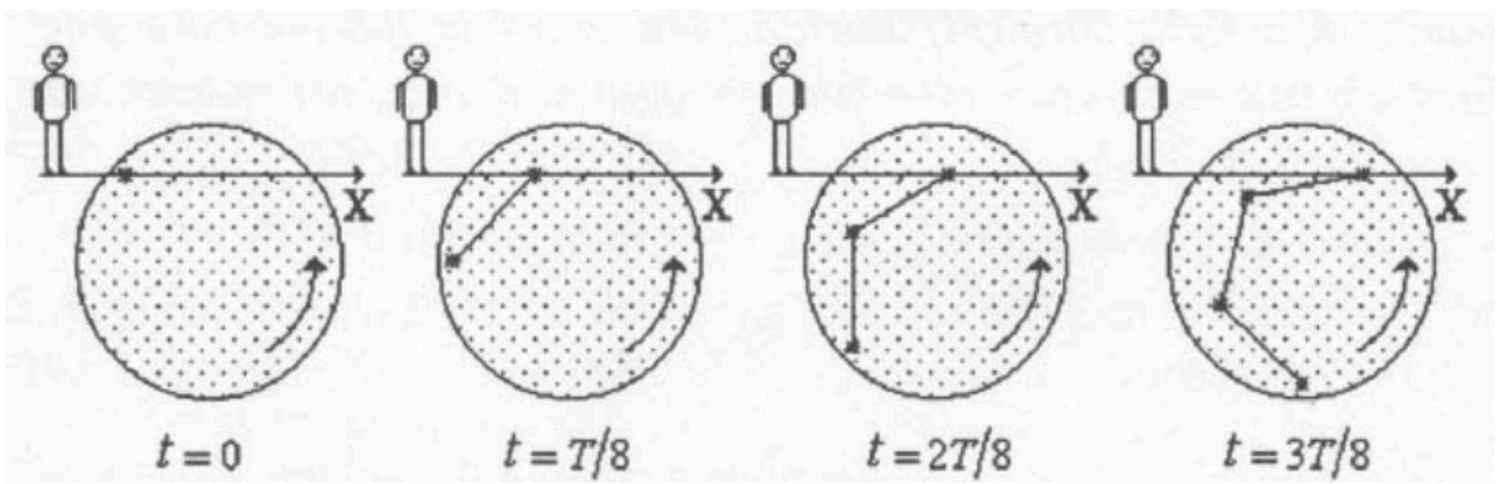


Fig **1.12** Motion relative to an inertial and a non-inertial observer. The stationary inertial observer sees the free object moving away from him in a straight line across the face of the rotating disc. However, suppose the disc rotates with period T and that at regular intervals, the object squirts a drop of ink onto it. As the pattern of ink spots shows, the path taken **by** the object as seen from the disc is curved and not straight. Thus an observer riding on the rotating disc would not ascribe a straight path to the motion of the object but a curved one, even though no actual force acts upon it.

This is the familiar *Coriolis effect.*

The atmospheric circulations - the cyclones and anticyclones - so important in interpreting terrestrial weather patterns, are the result of a Coriolis effect associated with the earth's rotation.

In order to account for these effects and to enable non-inertial observers to make sense of Newton's laws of motion, the existence of forces other than actual external forces must be postulated.

These *inertial* or *fictitious* forces act besides the actual external forces; it is they that push the passengers when their vehicle accelerates or brakes.

Actual forces are those exerted by one body on another, they are an expression of the interaction between the bodies.

An electrical interaction gives rise to an electrical force; magnetic interactions to magnetic forces, and so on.

The additional forces that appear in non-inertial systems are not the result of any interaction.

Their presence necessitates making modifications to the canonical form of the laws of motion that are particular to each non-inertial system; a modification valid in one non-inertial system will not necessarily work in another.

Thus, each non-inertial system has its own formulation of the laws of motion.

Evidently, Newton's laws of motion are truly covariant only for a special class of observers: inertial observers.

This usually presents no problem, as our mind's eye can always conjure up a hypothetical inertial observer, positioned at some convenient point in space, relative to whom the motion of the object in question can be analyzed using Newton's laws of motion.

Implicit in this approach is the assumption that all points in space are equivalent.

Newton expressed this idea through his postulate of an absolute space which "in its own nature, without regard to anything external, remains always familiar and immovable".

It is within this space that inertial observers can never be sure if it is they or the objects around them that are moving, and it was relative to this space that change in motion occurred.

In this way, unqualified use was made of the Newtonian conception of motion for some two hundred years.

However, with the development of electromagnetic theory, it was found to give rise to a profound paradox.

For it transpired that under this conception the laws of electromagnetism take a different form for different inertial observers, depending on whether they are at rest or in motion.

Thus, though stationary and moving inertial observers are indistinguishable by reference to Newton's laws, under the self-same conception of motion, they should be distinguishable by reference to laws of electromagnetism.

The following two case studies illustrate the problem.

Case 1: Coulomb's Law and Magnetic Forces

Consider the interaction between two identical positive charges when the observer is

(i) at rest relative to the charges;

(ii) in uniform rectilinear motion relative to the charges (Fig 1.13).

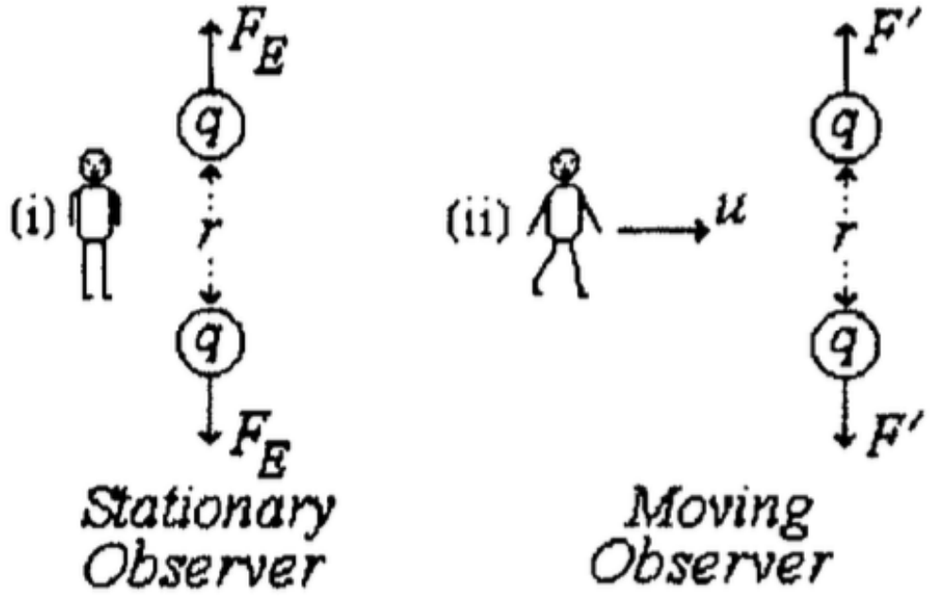


Fig 1.13 The forces acting between two positive charges as viewed by two inertial observers. Observer (i), who is at rest relative to the charges, detects the Coulombic force, F_E. However, observer (ii), who is in uniform motion relative to the charges, detects a weaker force F'.

When the observer is at rest relative to the charges, he finds that the magnitude of the repulsive force acting between them corresponds to that given by Coulomb's Law,

$$F_E = \frac{kq^2}{r^2}$$

However, when he is moving in a straight line with velocity u relative to the charges or they are moving similarly in parallel relative to him, he finds that a weaker force, $F' < F_E$, acts between them.

Law is seemingly not valid when the observer and the charges are moving relative to one another.

Thus, it should provide a basis for distinguishing between stationary and moving inertial observers.

Classical theory circumvents this difficulty by introducing a new phenomenon, that of *electromagnetism***.**

It postulates that by reason of the relative motion, a magnetic field B, is set up alongside the electric field.

This, in turn, bears upon the charges such that an attractive force of magnitude $F_M = F_E - F' = Bqu$ acts between them.

However, this really solves nothing, for it is only by virtue of the difference in the forces acting between the charges that we are made aware of the appearance of the magnetic field.

We have no other independent proof of its existence; no magnetic poles or magnetic charges have ever been isolated.

The magnetic field appears to be no more than an arbitrary contrivance - a mathematical fudge - that makes it possible to use Coulomb's law even when charges are in motion.

It has no independent reality.

Nor does its proposition really solve the fundamental problem of the apparent nonequivalence of the inertial observers with regard to Coulomb's law.

In order to establish their equivalence, Coulomb's Law must be shown to be valid for both inertial observers without the need for any special postulates, such as those made by electromagnetic theory, and which can themselves serve to distinguish between the observers.

The difference between the forces they measure must be explained in terms of a physical concept such as their relative motion, that has the same import for both.

This cannot be done under the Newtonian conception of motion.

Case 2: <u>Electromagnetic Induction</u>

The interaction between a magnet and a conductor moving at a uniform velocity, u**,** relative to one another, can be viewed from two points of view (Fig 1.14**):**

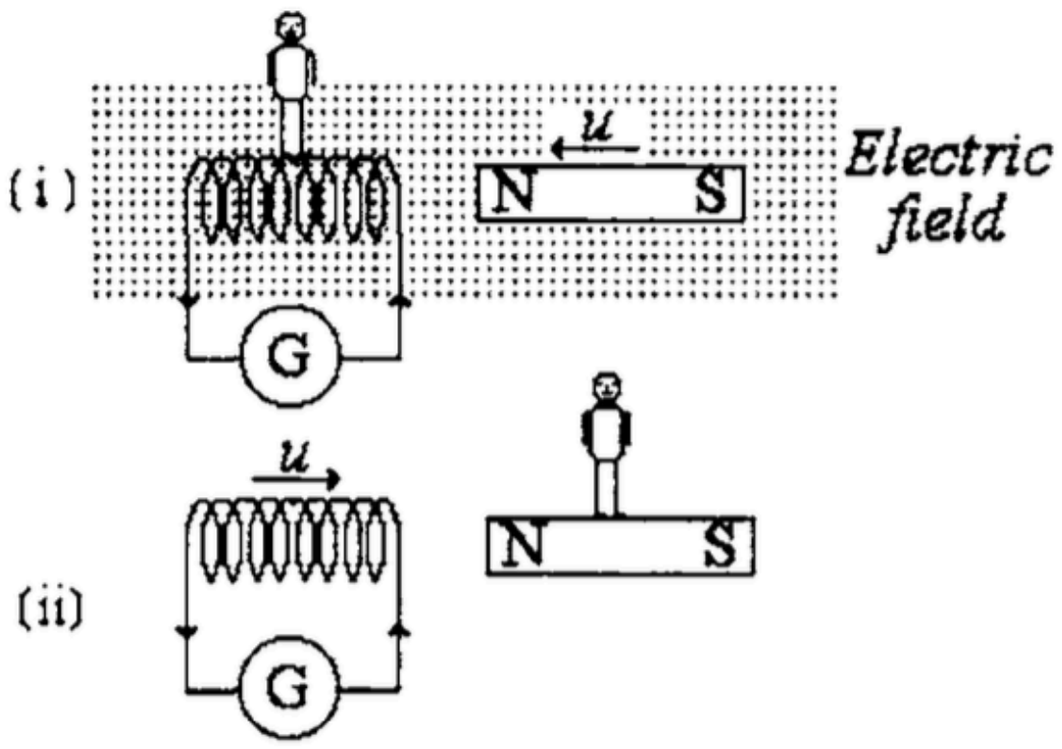


Fig 1.14 Electromagnetic induction viewed from different inertial systems. The coil and magnet are closing at a relative velocity u**.** Under the classical approach observer (i), at rest relative to the coil, should detect an electric field arising from the motion of the magnet, whereas observer (ii), at rest relative to the magnet, detects no electric field.

(i) from that of an observer at rest relative to a metal conductor, who sees the magnet approaching the conductor at a uniform velocity, *u*;

(ii) from that of an observer at rest relative to the magnet, who sees the metal conductor approaching the magnet at a uniform velocity, *u*.

Under the Newtonian conception of motion, the explanations that Maxwell's theory of electrodynamics gives for the two cases are different; it ascribes a different law to each of the inertial observers.

In the first case, it explains that the motion of the magnet sets up an electric field of a certain energy in the space around it which, in turn, excites an electric current in the stationary conductor.

In the second case, it states that an electromotive force, which of itself corresponds to no energy, is induced in the moving conductor as it moves through the stationary magnetic field; the electromotive force causes the electric current to flow.

However, experiment belies these distinctions.

In practice, electromagnetic induction experiments reveal no asymmetry between inertial observers; rest and uniform motion in a straight line are electromagnetically indistinguishable.

Just like the sailor on Newton's ship, the inertial observers cannot tell whether they or their surroundings are moving.

Einstein recognized that the problem lay with the application of the Newtonian conception of motion to the interpretation of electromagnetic phenomena.

Under the Newtonian rationale, magnetism cannot be shown to be just the effect of relative motion on electrical interactions nor can electromagnetic induction be given a single explanation valid for all inertial observers.

Despite its unprecedented achievements, Newtonian mechanics is in fact flawed.

The theory of special relativity arose from the need to find a new conception of motion which would answer these needs.

In order to accomplish this, Einstein had to overturn some of the most fundamental intuitive assumptions of classical physics, replacing them with new non-intuitive concepts that challenge our normal experiences.

Thus, our study of the theory of relativity must start with a re-examination, in some depth, of the assumptions underlying the Newtonian conception of motion.

1.3.2 The Newtonian Conception of Motion

The science of mechanics is concerned with the motion of bodies or objects and the effect of forces upon them.

It defines motion as change in position; a moving body is one whose position is changing.

Classical (Newtonian) mechanics describes the motion of an object in terms of its trajectory through space.

This is the continuous line obtained by plotting the object's instantaneous positions as it moves through space**.**

Positions in space are most usefully specified by means of a coordinate system in which every point in space is defined by its unique set of coordinates.

The coordinate system constitutes the ***reference frame*** (or ***frame of reference)*** relative to which measurements of position are made (Fig 1.15).

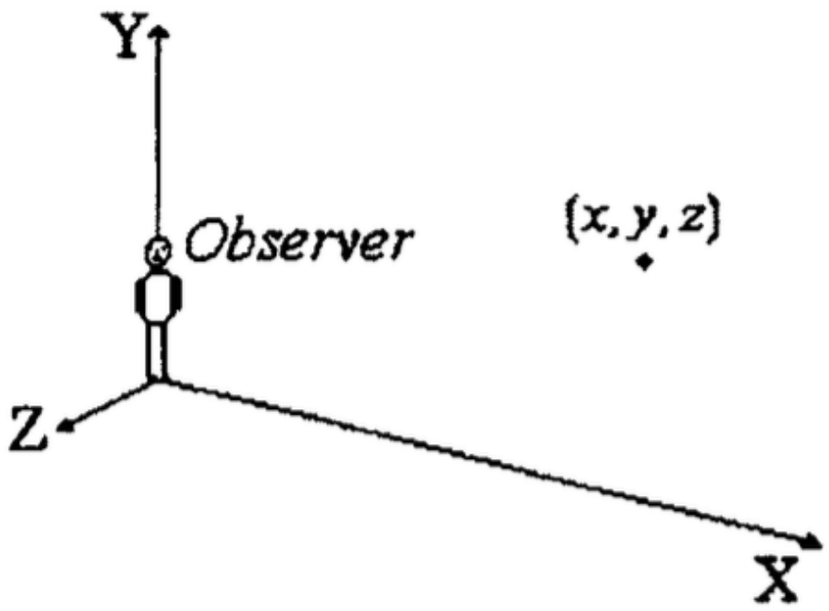


Fig 1.15 An observer and the orthogonal coordinate system which is his chosen reference frame. Each point in space is defined by a unique set of coordinates **(x, *y, z)***

Reference frames *can* be at rest, in uniform motion in a straight line or in an accelerated motion.

Reference frames at rest or in uniform motion in a straight line are called inertial reference frames.

However, notwithstanding their fundamental equivalence, because of their different reference frames, each of the observers may well see things differently; they will not necessarily record identical observations and measurements even when studying the same phenomenon.

How can their different sets of observations and measurements be shown to be essentially equivalent?

How can the laws different observers formulate be shown to be identical?

How, in fact, can the same law arise from different sets of data?

Consider two inertial observers, S and ***S'***, whose chosen reference frames are in uniform motion relative to one another (Fig 1.16).

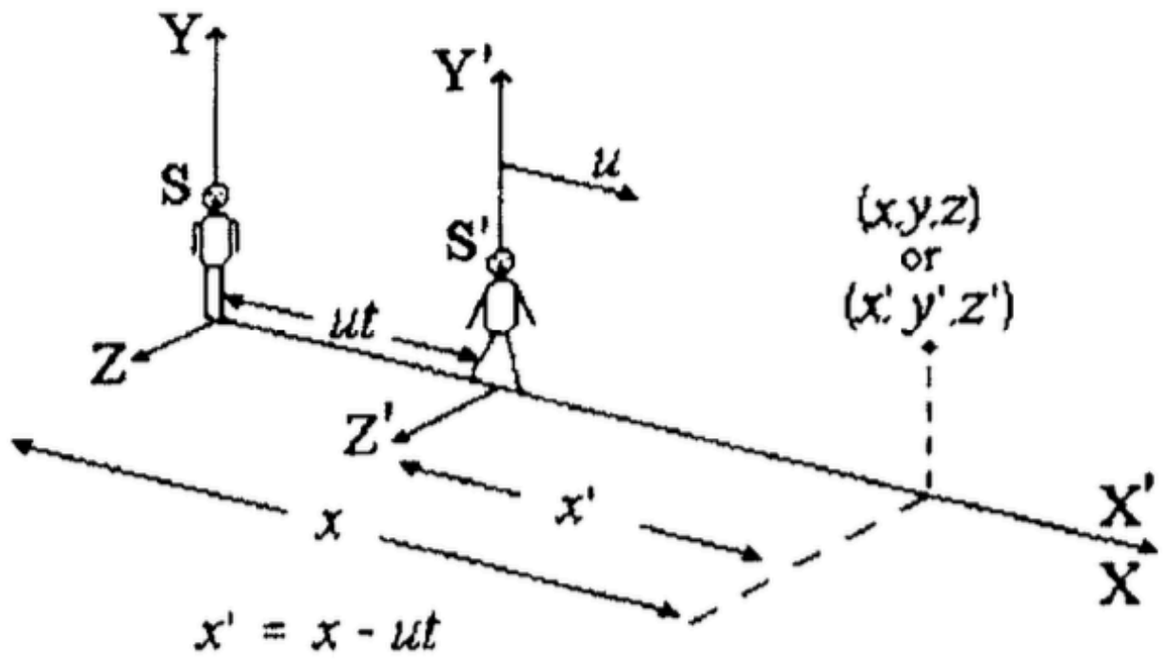
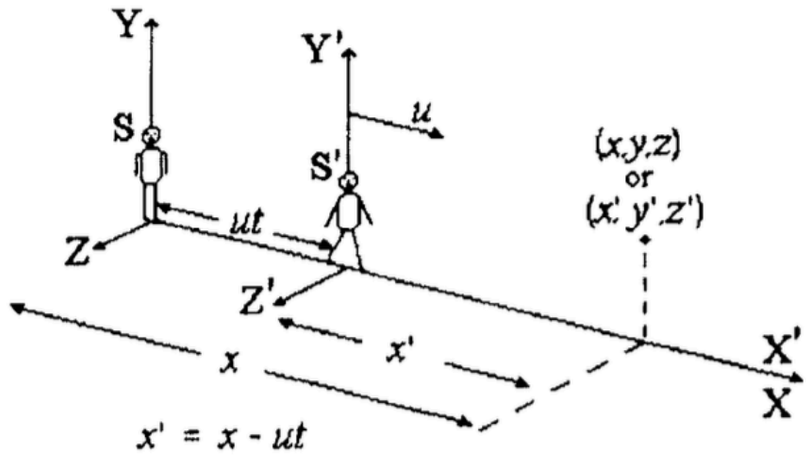


Fig 1.16 Observer S' is moving along the X axis with velocity u relative to the observer S. An event that occurs in their common space is seen by both observers. Observer S records that the event took place at the point whose coordinates relative to him are(x, *y*, z); S' records that it took place at the point (x', *y*', z').

Observer S' might be one of the passengers in a bus traveling at a constant velocity, u, along a flat straight road, the XX' axis; observer S might be a person standing at the road-side.



Though both are inertial observers, the passenger sitting in the bus, S', sees many things differently from the observer S standing at the road-side.

For example, a point whose coordinates relative to S' are (x',y', z') has a different set of coordinates, (x, *y*, z), relative to the observer S.

Moreover, whereas the passenger S' sees all the other passengers sitting in the bus as being at rest, relative to observer S standing at the road-side, all the bus passengers including ***S***' are in motion, travelling with the same velocity, u, as the bus.

How can these evidently different observations be shown to be equivalent and to result in the same physical law?

In order to establish (or refute) the underlying equivalence of the sets of observations and measurements made from different reference frames and of the physical laws derived from them, a systematic independent procedure is required.

In the language of the mathematicians, such a procedure is called a transformation.

If the application of the transformation to the observations and measurements made from one reference frame systematically converts them all into the corresponding observations and measurements made from a different reference frame, then the two sets of observations and measurements are essentially equivalent.

If it doesn't, then they are not.

The same applies to the laws formulated by the observers in each reference frame.

They are covariant laws if the application of the transformation to the formulation made by one observer produces the formulation made by another; the laws are then said to be *form invariant*.

The laws of physics (and of the other 'exact' sciences) are typically formulated in terms of measurable physical quantities.

In general, for physical laws to be covariant, the same transformation that reconciles measurements of the physical quantities made by one inertial observer with those made by another, must also convert the laws formulated by the one into those formulated by the other.

The transformation appropriate to the Newtonian conception of motion is called the ***Galilean Transformation*** and, in the case where the relative motion of the reference frames is parallel to the X-axis, it can be derived as follows.

Suppose the two observers, S and S', possess identical watches, which they synchronize and zero at the moment they pass one another.

Some time later, the observer S notices the occurrence of an event and records that it took place at a time t according to his watch and at a point in space whose coordinates, relative to him, are (x, *y, z)*.

The second observer, S', who is traveling along the XX' axis, notices the same event and records that it took place at a time t' according to his watch and at a point in space whose coordinates, relative to him, are (x', *y'*, *z'*) .

What is the mathematical relationship between the space and time coordinates, x, y, z and t, recorded by observer S and those, x', y', z' and t', recorded by the observer S'?

In particular, what is the relationship between t and t'?

The classical answer to this question was given by Newton in 1687 in the framework of the mechanical system described in his great work ***Principia.***

He wrote:

> *Absolute, true, and mathematical time, of itself and from its own nature, flows equably without relation to anything external… All motions may be accelerated and retarded, but the flowing of absolute time is not liable to any change … the existence of things remains the same, whether the motions are swift or slow, or none at all.*

In Newton's world, time is innate and absolute: an inherent property, like the mass of an object, affected neither by motions nor events.

In this world, "time is the medium of narration" and *past*, *future*, *present* and *simultaneous* have clear and simple meanings that correspond to our intuitive understanding of them.

The past comprises those events of which we may have knowledge but over which we have no influence.

The future will be filled by events over which we may have some influence but of which we have no present knowledge.

The present is just a passing moment common to all observers, irrespective of their positions or motions; an infinitesimal that separates past events from future ones.

And simultaneity is an inherent property of events themselves; events cannot be simultaneous relative to one observer but not **so** relative to another.

In the Newtonian conception, the passage of time is an invariant quantity.

Hence, under the Newtonian conception, using their synchronized watches, the two observers, S and S', must record that the event they both saw occurred at the same time, i.e.

$$t' = t \tag{1.9}$$

Although the two observers, S and S', assign identical time coordinates to the event, the same is not necessarily true of the space coordinates; these are *relative* measures whose values do depend on the reference frame from which they are taken.

Because of their relative motion along the XX' axis, x coordinates recorded by the observer S will be greater than the corresponding x' coordinates recorded by the observer S' by an amount ut.

$$x' = x - ut \tag{1.10}$$

There being no relative motion parallel to the YY' or ZZ' directions, both observers record the same coordinate values in these directions.

$$y' = y \tag{1.11}$$

$$z' = z \tag{1.12}$$

These four equations, (1.9), (1.l0), (1.11) and (1.12), constitute the ***Galilean Transformation.***

They provide the means for the systematic conversion of the space and time coordinates determined by either of the inertial observers, S or S', into the corresponding coordinates specified by the other.

In this way, the underlying equivalence of their observations and measurements can be validated.

An important corollary of the Galilean Transformation, is that although inertial observers like S and S' do not assign the same coordinates to a position in space, they will always agree as to the distance between two points in space or as to an object's size and dimensions.

Consider the rod AB in Fig 1.17.

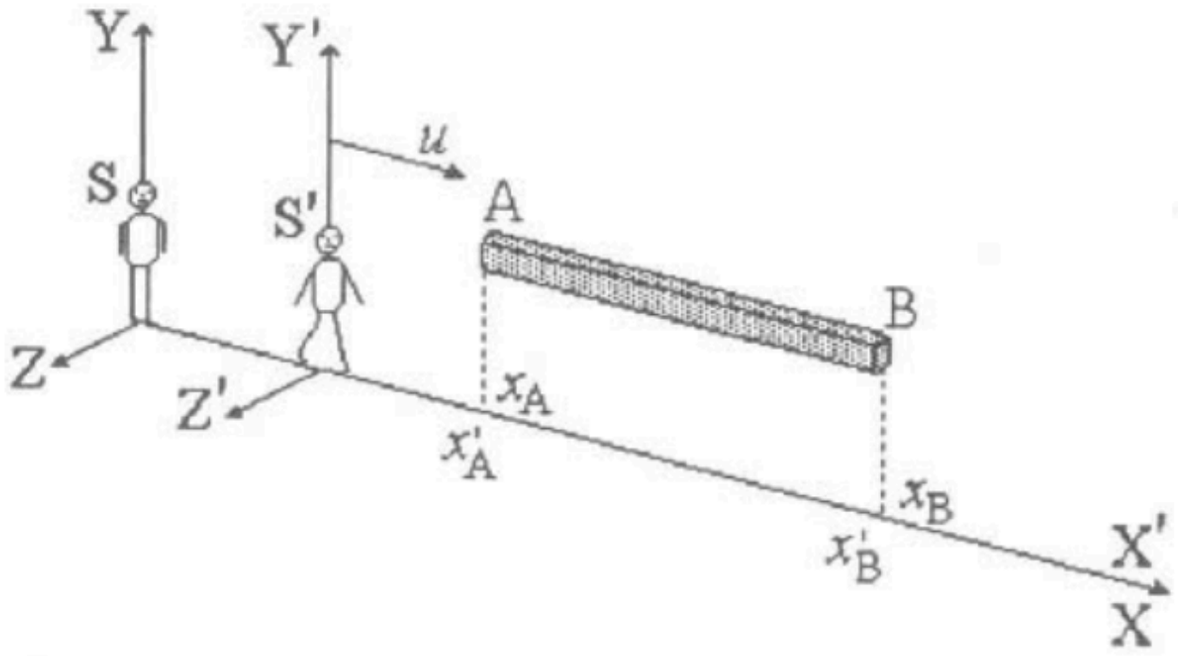


Fig 1.17 Two points, A and B, marking the ends of a rod in the stationary reference frame S**.** The length of the rod according to the coordinate system of S is $x_B - x_A$ and according to the coordinate system S' is $x'_B - x'_A$.

Relative to observer S, the length of the rod is $x_B - x_A$, whereas relative to ***S'*** it is $x'_B - x'_A$.

However, from equation (1.10)

$$\begin{aligned} x'_B - x'_A &= (x_B - ut) - (x_A - ut) \\ &= x_B - x_A \end{aligned} \tag{1.13}$$

Thus, in the Newtonian conception, the distance between two points (the length of the rod) is an invariant quantity whose value is unaffected by the motion of the inertial reference frame from which it is measured.

Now suppose, that the two observers, S and ***S'***, measure the velocity and acceleration of a moving object.

What is the relationship, under the Newtonian conception, between the object's velocity, v, as measured by the observer S, and its velocity, v', as measured by the observer S'?

Likewise, what is the relationship between its acceleration, **a**, relative to S and that, **a'**, measured by the observer S'?

The respective scalar components, v_x, v_y, v_z, and v'_x, v'_y, v'_z, of the object's velocity that the two observers, S and S', record are obtained by taking the derivatives of equations (1.l0), (1.11) and (1.12) of the Galilean transformation with respect to time.

$$\frac{dx'}{dt} = v'_{x'} = v_x - u \tag{1.14a}$$

$$\frac{dy'}{dy} = v'_{y'} = v_y \tag{1.14b}$$

$$\frac{dz'}{dt} = v'_{z'} = v_z \tag{1.14c}$$

These three equations *can* be combined into the single vector equation,

$$\mathbf{v'} = \mathbf{v} - \mathbf{u} \tag{1.15}$$

known as the *velocity addition rule of the Galilean Transformation*, which, like all vector relationships, carries its transformation within it.

A corollary of the velocity addition rule is that whenever two observers are moving relative to one another, i.e., $\mathbf{u} \neq 0$, the velocities, **v** and **v'**, that they respectively assign to an object will be different.

Thus, like its position, on object's velocity is a relative quantity, i.e., one whose value depends on the reference frame from which it is measured.

To express an object's velocity or position without indicating the reference frame from which it was measured, is to make a meaningless statement.

The relationship between the accelerations measured by the two inertial observers, S and S', is obtained by taking the derivative of equation (1.15) relative to time.

Since their relative velocity, **u**, is constant, this gives:

$$\frac{d\mathbf{v}'}{dt} = \mathbf{a}' = \mathbf{a} \tag{1.16}$$

Thus, observers in different inertial reference frames will assign the same value to the acceleration of the object.

Under the Galilean Transformation, acceleration is an invariant quantity.

Given that mass too is an invariant, the invariance of force follows from Newton's second law of motion, **F = ma.**

Returning to the passengers in the bus and the observer at the roadside, we can now reconcile their apparently conflicting observations:

* **Relative to the bus passenger, S′, the velocity of all the other bus passengers equals zero, i.e., $\mathbf{v}' = 0$.**

* **Relative to the observer at the road-side, S, the passengers' velocity is $\mathbf{v} = \mathbf{u}$, where $\mathbf{u}$ is the velocity of the bus.**

* **However, substituting these two velocities, $\mathbf{v} = \mathbf{u}$ and $\mathbf{u} = \mathbf{u}$, in the transformation equation (1.15) gives for the velocity, $\mathbf{v}'$:**

$$\begin{aligned} \mathbf{v}' &= \mathbf{v} - \mathbf{u} \\ &= \mathbf{u} - \mathbf{u} = 0 \end{aligned} \tag{1.17}$$

which is just what the passenger S′ determines. Thus the *conflicting observations* are reconciled; they are shown to be essentially equivalent under the Galilean Transformation.

To summarize the Newtonian conception of motion:

1. **Measurements of position or velocity made from different inertial reference frames such as S and S′ will generally differ; position and velocity are relative measures. However, measurements of position and velocity made from one inertial frame can be reconciled with those made from another by means of the Galilean Transformation.**

2. Measurements of distance (length), time, mass and acceleration made from different inertial reference frames will be identical; distance, time, mass and acceleration are invariant quantities.

3. All inertial observers formulate the same laws of mechanics, that are covariant under the Galilean Transformation. In order to convert a law discovered by the observer S′, into the form it must have for the observer S, the appropriate transformation equations for the relative quantities, x', y', z' and v', in terms of x, y, z and v are substituted in the formulation of the law made by S′.
4. In general, the Newtonian conception of motion corresponds to our intuitive conception of motion. It harbours no surprises.

1.3.3 The Michelson-Morley Experiment

Although Maxwell's electromagnetic theory does not specify any particular or preferred reference frame for the velocity of light in vacuo, the existence of the aether was so deeply rooted in 19th century physics that it was generally presumed that it must be the reference frame for the velocity of light in vacuo.

Just as the air is quite naturally taken to be the reference frame for the velocity of sound, it was simply taken for granted that the aether was the reference frame for the velocity of light.

It was likewise assumed that the propagation of light conformed to the Newtonian conception of motion.

Thus, the velocity of light was presumed to be a relative quantity subject to the same velocity addition rule, viz. $\mathbf{v}' = \mathbf{v} - \mathbf{u}$, as the velocity of sound.

On this assumption, the velocity of light relative to a reference frame other than the aether should be given by **c' = c - u,** where **c** is its velocity relative to the tether and **u** is the velocity of the reference frame relative to the aether.

For example, an observer moving towards a stationary light source, or one towards whom the light source is moving, should measure a light speed greater than ***c.***

Similarly, an observer moving away from the source, or one from whom the source is receding, should measure a speed less than ***c.***

If their relative speed is **u,** the former observer should find the speed of light to be ***c + u*** and the latter should find it to be ***c*** - ***u***.

Furthermore, should the light appear to the observer to be propagating along a path at right angles to his own, its speed relative to him should be $\sqrt{c^2 - u^2}$.

The validity of these assumptions was investigated experimentally by Michelson and Morley in 1881.

Their experiment set out to detect the effect of the motion of an observer through the tether on measurements he made of the speed of light in *vucuo***.**

The motion was provided by the earth itself, which was known from astronomical calculations to be orbiting the sun at a speed of approximately 30,000 m/s (Fig 1.19).

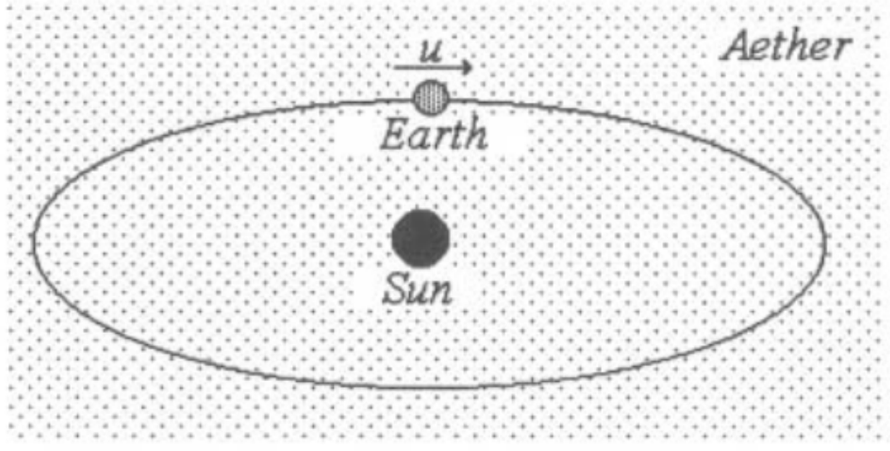


Fig 1.19 The orbital motion of the earth around the sun through the supposed all pervading ether.

It was presumed that the earth passes through the all-pervading aether in its orbital motion around the sun.

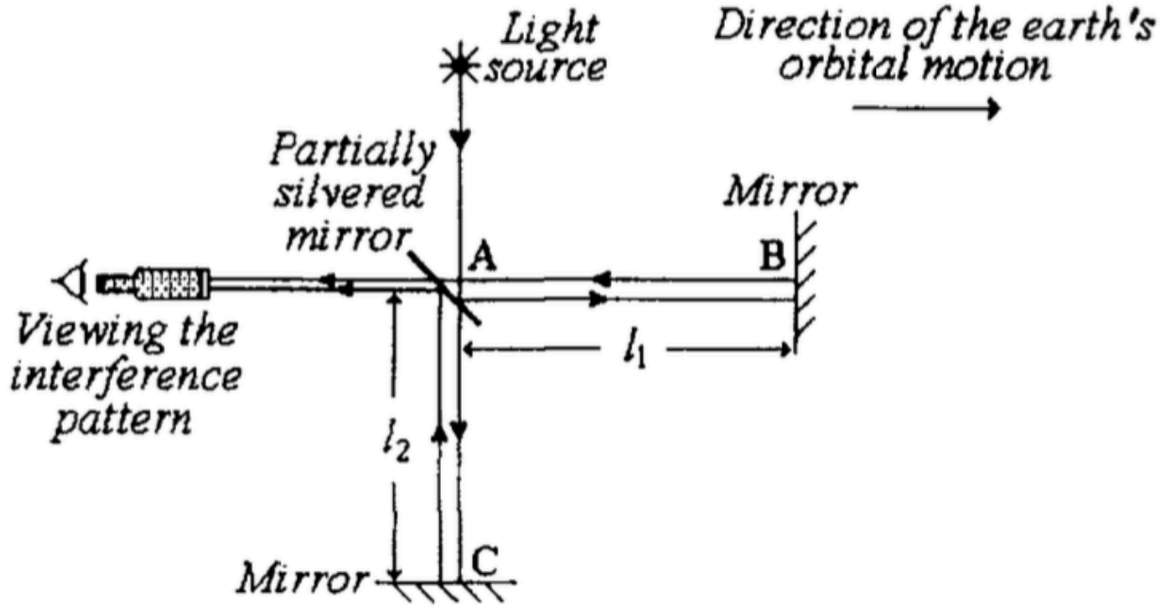
Accordingly, just as the motion of an observer through still air affects his measurements of the speed of sound, the earth's motion through the still aether was expected to affect terrestrial measurements of the speed of light.

The orbital speed of the earth is four orders of magnitude smaller than the speed of light in vacuo, c = 300,000,000m/s**,** and so the maximum difference that Michelson and Morley could expect to find in their terrestrial measurements of the speed of light was about 0.0002c**.**

The detection of such a small difference in the speed of light required great accuracy and precision in the design and execution of the experiment.

For this purpose, Michelson perfected a new and very accurate instrument-the optical interferometer (Fig 1.20).

Fig 1.20 The Michelson-Morley experiment. A light beam from a monochromatic source is split into two coherent beams by the partially silvered mirror. After reflection from the mirrors, the beams are recombined and the interference pattern produced is observed by means of a suitable telescope. The apparatus was mounted on a heavy rotatable horizontal base so that it could be set at different orientations relative to the direction of the earth's orbital motion.



As its name implies, an optical interferometer is a measuring device that works on the basis of the interference patterns produced by visible light.

In a typical instrument, monochromatic light is first split into two coherent beams.

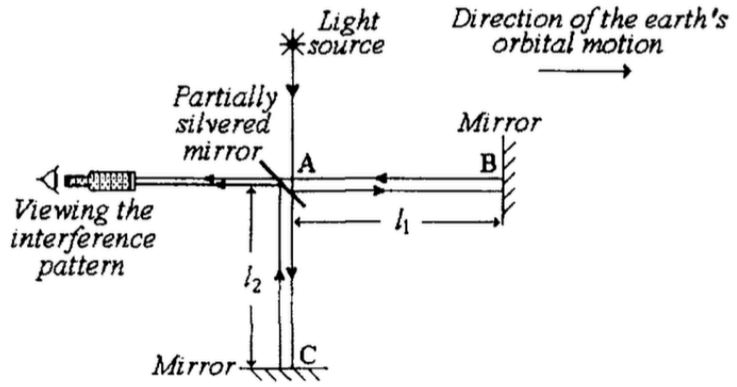
After traveling along different paths, the two beams are recombined producing an interference pattern.

Where precisely the bright and dark fringes appear in the interference pattern depends on the phase difference between the recombined beams.

This, in turn, depends on two factors: (i) the difference between their respective path lengths and (ii) the difference between their travel times.

Any change in the phase difference between the beams will show up as a shift in the position of the interference fringes.

In the Michelson and Morley experiment, the whole apparatus was mounted on a heavy rotatable horizontal base so that observations could be made at different orientations relative to the direction of the earth's orbital motion.



The paths of the two light beams, ABA and ACA, were set at right angles to one another and their lengths, l_1 and l_2 , respectively, were fixed.

They remained so throughout the experiment.

Consequently, any shift in the position of the interference fringes as the orientation of the apparatus was changed, could only be the result of variations in the difference, $\Delta t = t_{ABA} - t_{ACA}$, between the beams' travel times.

The travel time of each path depended in turn on the speed relative to the terrestrial observer, c', at which the light propagated along it.

Thus, any shift in the position of the fringes would have revealed that the speed of light varied with the direction of its propagation relative to the earth's motion and, by inference, would have confirmed the aether's existence.

Suppose the apparatus is initially set up as shown in Fig 1.19, i.e., with the ABA beam parallel to the direction of the earth's motion and the ACA beam at right angles to it.

According to the Newtonian conception, as the light travels through the still Ether from A to B, its speed relative to the terrestrial viewer is ***c - u;*** as it returns to A, after reflection from the mirror B, its speed relative to this same observer is *c* + *u***.**

As regards the other beam, ACA, the terrestrial viewer sees it as propagating along a path perpendicular to his own and so its speed relative to this observer is $\sqrt{c^2 - u^2}$.

Hence, the difference, Δt, between the beams' travel times should be

$$\begin{aligned}\Delta t &= (t_{AB} + t_{BA}) - (t_{AC} + t_{CA}) \\ &= \left(\frac{l_1}{c-u} + \frac{l_1}{c+u}\right) - \left(\frac{l_2}{\sqrt{c^2-u^2}} + \frac{l_2}{\sqrt{c^2-u^2}}\right)\end{aligned} \tag{1.18}$$

Suppose the apparatus is now rotated through 90° so that the ACA beam turns parallel to the direction of the earth's motion and the ABA beam is now at right angles to it.

The new value for the difference between the beams' travel times should be

$$\begin{aligned}\Delta t' &= (t_{AB} + t_{BA}) - (t_{AC} + t_{CA}) \\ &= \left(\frac{l_1}{\sqrt{c^2-u^2}} + \frac{l_1}{\sqrt{c^2-u^2}}\right) - \left(\frac{l_2}{c-u} + \frac{l_2}{c+u}\right)\end{aligned} \tag{1.19}$$

Clearly, $\Delta t \neq \Delta t'$ and so the rotation should result in a shift of the fringes in the interference pattern.

Michelson and Morley made repeated observations of the interference pattern with their apparatus set each time in a different orientation relative to the earth's orbital motion.

However, in not a single case was a significant shift observed.

The implications of this null result were awesome.

Evidently, terrestrial measurements of the speed of light do not depend on the direction of the light's propagation relative to the earth's motion; the propagation of light is not analogous to the propagation of sound there is no luminiferous aether; the earth does not move through the aether as it orbits the sun; and light waves are not aether waves.

In an attempt to 'save the aether' and to account for the null result of the Michelson and Morley experiment within the framework of the Newtonian conception of motion and the Galilean transformation, it was suggested that the aether is somehow dragged along with the earth in its motion through space, such that their relative velocity is zero.

Experiments were carried out in an attempt to detect this *aether drag* but they were all inconclusive.

Thus, the unavoidable conclusion was that the aether did not exist.

This, however, posed a host of new questions:

(i) If light waves are not æther waves, then just what sort of waves are they? How are we to understand the physical nature of light and of electromagnetic waves? Whereas sound waves require a material medium to carry their energy, electromagnetic waves appear to propagate without the need for any detectable medium. How can a wave exist and carry energy through absolutely empty space? What reality can we ascribe to a wave which propagates without the need for any physical medium at all?

(ii) Velocities are relative quantities; they are meaningless unless related to a reference frame. If the æther does not exist, what is the reference frame for the velocity of light *in vacuo*?

(iii) If, as the null result of the Michelson and Morley experiment indicates, the Galilean Transformation does not correctly reconcile measurements of the velocity of light *in vacuo* from different inertial reference frames, how exactly is it flawed?

(iv) Is the flaw in the Galilean transformation specific to the velocity of light or is it a fundamental flaw that affects its application to all phenomena, both mechanical and electromagnetic?

(v) If it is a fundamental flaw, why does the transformation still work so well in our everyday experience?

(vi) If the Galilean transformation is indeed fundamentally flawed, what is the correct transformation to use when reconciling observations made in different inertial reference systems?

All of these questions, except for that concerning the physical nature of light, were first answered by Einstein's *Special Theory of Relativity*.

The question of the actual physical nature of light would have to await the development of the quantum theory of light.

1.3.4 The Postulates of the Special Theory of Relativity

The Special Theory of Relativity is derived from two postulates:

1. The Principle of Special Relativity

The laws of physics have the same form in all inertial reference frames

The Newtonian principle of relativity asserts that the laws of mechanics shall have the same form in all inertial reference frames.

Einstein widened the scope of this principle to include all the laws of physics.

He asserted that all the laws of physics, without exception, must be covariant.

This extension of the Newtonian principle of relativity recognizes the fact that no experiment, either mechanical or electromagnetic, has ever succeeded in determining whether an inertial reference frame is in absolute motion or at absolute rest.

Experimentally, all inertial frames are indistinguishable.

The laws of physics must acknowledge this experience.

2. <u>The Principle of of the Constancy of the Speed of Light</u>

The speed of light *in vacuo, c,* has the same value in all directions in all inertial systems.

Einstein's solution to the problem of the relativity of the speed of light was simple but shattering, for it challenged all our intuitive conceptions of space and time.

Taking the results of the Michelson-Morley experiment at their face value, he simply declared that the speed of light in *vacuo,* c *,* has the same value relative to every inertial system, i.e., it is an invariant quantity.

In effect, Einstein raised the speed of light in vacuo to the status of a law of nature, true for all inertial observers.

Consequently, instead of relating the speed of light in *vacuo* to the reference frame of the aether, it can be related to any inertial frame we choose; its value will always be c relative to the frame chosen.

Thus, there is no need for the aether.

Every inertial frame is its own reference frame for the velocity of the light in *vacuo.*

Whether the inertial frame is moving towards the light source or away from it, measurements of the speed of light in *vacuo* made from the frame will all give the same value ***c*.**

Since the original Michelson and Morley experiment, the constancy of the speed of light in *vacuo* in inertial systems has been directly confirmed many times.

In one particularly precise experiment carried out in 1964 at the CERN particle-physics laboratory, the speed of the high frequency electromagnetic waves (γ rays) emitted from the decay of unstable short-lived particles called neutral pions was measured.

A high-velocity, u = 0.99975c**,** beam of neutral pions was produced in a particle accelerator.

The researchers measured the speed relative to the laboratory of the electromagnetic waves emitted from these high-velocity particles as they decayed.

The value they obtained was $c' = 2.9977 \times 10^8$ m/s as compared to the accepted value for the speed of light in vacuo of $c = 2.9979 \times 10^8$ m/s .

Plainly, the high velocity of their source, close itself to the speed of light, had no effect on the speed at which the electromagnetic waves propagated relative to the laboratory; had the source been stationary the same value would have been measured.

The experiment was clear confirmation that $c' = c$ even if $u \neq 0$.

1.3.5 Simultaneity and the Relativity of Time

In developing his mechanical system of the world, Newton had postulated two absolutes: *absolute space* and *absolute time*.

These were the pillars upon which the Newtonian conception of motion and its attendant Galilean transformation were established.

However, as we have seen, this edifice is flawed; it does not pass the test of covariance.

To find out why, we must re-examine the validity of Newton's original postulates.

In **particular,** we must ask whether Newton's assertion of the absolute nature of time is compatible with the proven invariance of the speed of light?

Our enquiry will be conducted through a series of *thought experiments*.

<u>Thought Experiment 1</u>

Imagine two spaceships, S and S′**,** somewhere deep in space.

S is at rest and S′ is traveling in a straight line with a constant velocity u relative to S; the two spaceships are inertial reference systems (Fig 1.21a).

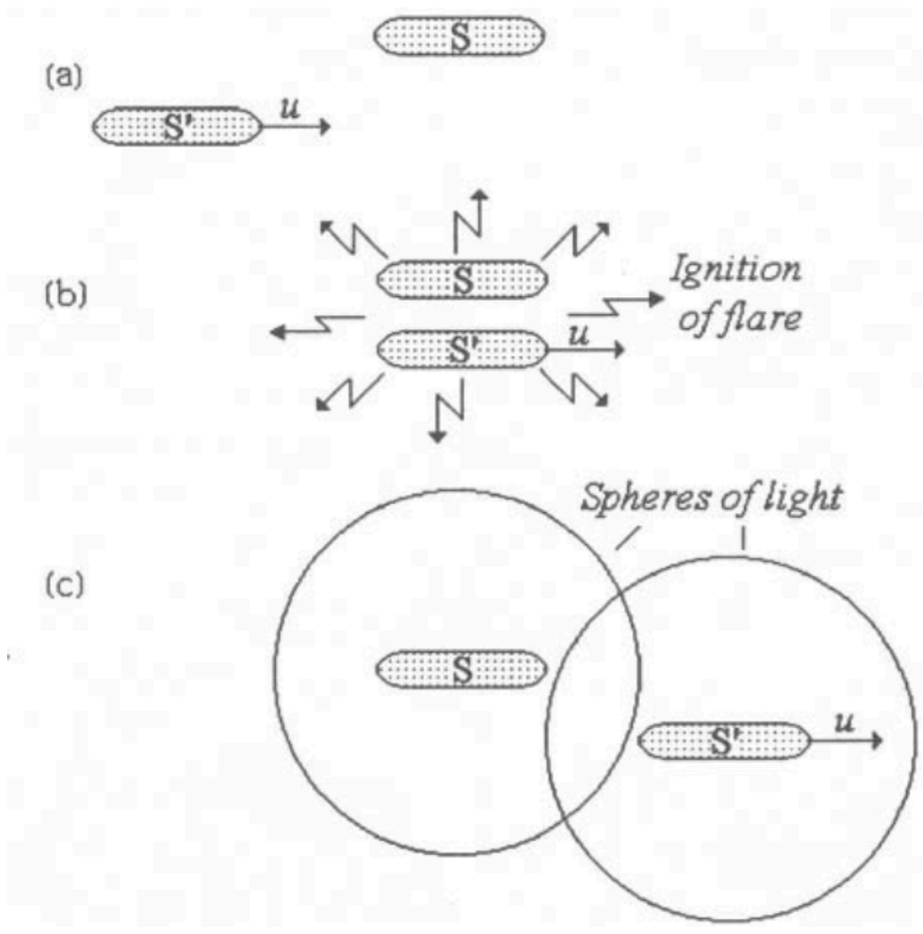


Fig 1.21 An encounter between two spaceships. (a) S is at rest as S′ approaches in a straight line with a constant velocity ***u.*** (b) As S′ passes S a flare is ignited. (c) Observers on each of the two spaceships see a sphere of light, reflected by the inter-stellar dust, radiating out from them at its center. For purposes of visualizing the phenomenon two spheres of light are shown in the figure though in reality no observer of the event would see more than one.

Suppose, that at the moment spaceship S′ passes spaceship S, a flare is ignited from one of them.

Its light spreads out into space uniformly in all directions (Fig 1.21b).

After the firing of the flare, the spaceships separate, S′ moving on away from S at the constant velocity u.

As the light spreads out into space it is scattered by the fine inter-stellar dust.

In accordance with the principle of the constancy of the speed of light, observers on each spaceship see a sphere of scattered light expanding away from them in all directions at speed c.

The same light is seen by the observers on both spaceships, for only one flare was ignited.

However, each observer sees himself standing at the centre of the sphere of light; despite the fact that the spaceships are moving apart at a relative velocity u, observers on both of them see identical phenomena (Fig 1.21c).

How is this possible?

How can the light from the flare appear in exactly the same form to the two observers, S and S′,

In order to elucidate this paradox, we will follow in Einstein's footsteps and start, as he did in his celebrated article on special relativity, by reconsidering the very nature of time and space and the methods of measuring them. Einstein wrote:

> *I f an observer at point* A *has a clock, he can estimate the time* of *events occurring in the immediate neighborhood of* A*, by looking for the position of the hands of the clock that is synchronous with the event. If an observer at point B possesses an identical clock to* A, B *he can similarly estimate the time* ***of*** *events occurring at B. But without further premises, it is not possible to compare, as far as time is concerned, the events at B with those at* A*. We have therefore an A time and a B time, but no time common to A and B.*

Einstein asserted that in order to compare events at the two points we must first establish a time common to both of them, i.e., we must first synchronize the clocks at A and B.

On this view, time is not a natural quantity, that "flows equably without relation to anything external" and that is present in nature whether measured or not.

On the contrary, it is a 'manufactured' quantity that depends on how and where it is defined.

The assertion that time and synchronization are constructs and not inherent properties of the physical world is the underlying notion of the Einsteinian revolution.

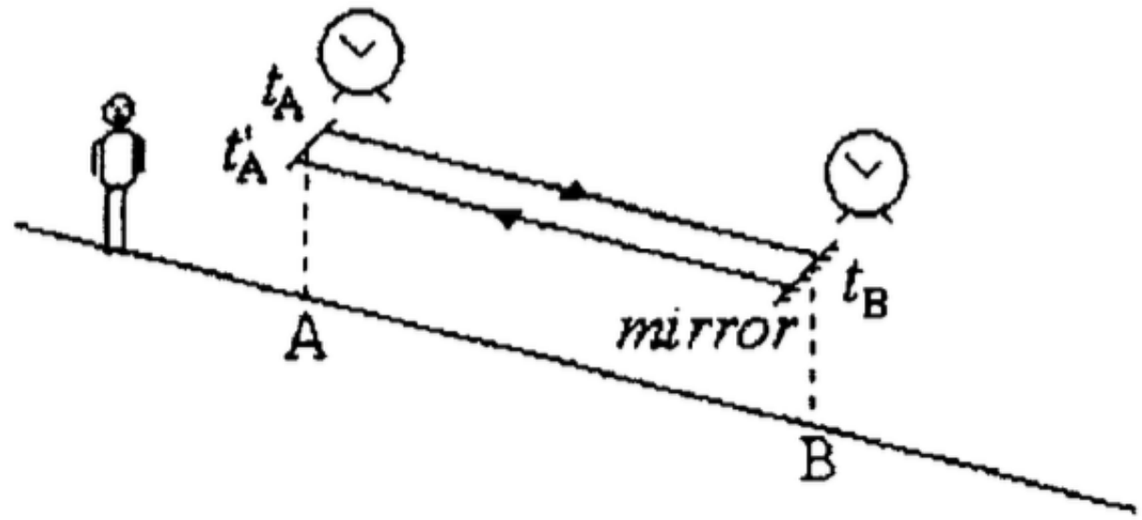
If synchronism is not automatic and inherent, its existence must be experimentally confirmed in each case.

Einstein argued that the synchronization of clocks situated at the two different points, A and B, in the same inertial system could be established as follows:

> A *common time can be defined [for any inertial system] if we establish by definition that the time which light requires to travel from A to B is equivalent to the time required* for *it to travel from B to A (Fig 1.22). For example, a ray of light proceeds from* A *at* A *time* t_A *towards B, arrives and is reflected from B at B time* t_B *and returns to A at A time* t'_A. *According to the definition, both clocks are synchronous, if*

$$t_B - t_A = t'_A - t_B \tag{1.20}$$

Fig 1.22 Clocks at rest in the same inertial reference system, S. A ray of light proceeds from A at A time t_A towards B, arrives and is reflected from B at B time t_B and returns to A at A time t'_A. The clocks are synchronous if $t_B - t_A = t'_A - t'_B$, i.e., if according to them the light takes the same time to travel from A to B as from B to A.



Implicit in this definition is the assumption that the speed of light from B to A equals that from A to B, as required by the principle of the constancy of the speed of light. Einstein continued:

> *Thus with the help of certain physical experiences, we have established what we mean when we speak of clocks at rest at different points in space and synchronized with one another; and thereby we have arrived at a definition of synchronism and time*.

Having established that the clocks are synchronous, they can now be used to determine whether events occurring at A and at B are simultaneous.

They can also be moved to other places in the reference frame so as to check whether events occurring at them are simultaneous or not.

The next step is to examine whether clocks situated in different inertia1 reference frames can also be synchronized.

For example, can a clock in the S inertial system be synchronized with one in the S' inertial system such that a common time can be defined for the two reference frames?

Were it possible to signal instantaneously, i.e., to transmit a signal at an infinite speed**,** we could synchronize the clocks without difficulty, irrespective of the relative motion of S and ***S'*****.**

An observer in one of the systems would simply signal the time shown by his clock to an observer in the other system.

The signal would be received instantaneously by the second observer who would set the time on his own clock accordingly**.**

But no method of signaling faster than flashing a light has ever been discovered and the speed of light in *vacuo***,** though extremely fast, is finite; it also has the same value in all inertial reference frames, irrespective of their motion.

Thus, signals cannot be instantaneously transmitted.

Classical Newtonian mechanics ignores this simple fact.

It acknowledges no upper limit on the speed at which objects or signals can move.

Thus it implicitly assumes instantaneous signaling is possible.

Consequently, it can posit the synchronization of all clocks and the existence of an absolute time common to all observers.

This is the essential flaw in classical mechanics.

In order to illustrate the problem of finding a common time for two different inertial systems and of determining the simultaneity of events occurring in them, we will carry out two further thought experiments.

Thought Experiment 2

Suppose, that flashes of light from two different light sources, each situated at the same distance, *d,* from an inertial observer, are seen by him at the same moment.

Since the sources are equidistant from him and the speed of light in an inertial system has the same value c in all directions, the observer must conclude that bursts of light were emitted by the sources simultaneously and at a time t = d/c before he saw them.

Whatever his motion, there is no way that the light from one source could have taken longer to reach the inertial observer than that from the other source.

Thought Experiment 3

Fig 1.23 shows two spaceships, that serve as inertial reference systems for two observers, Jack (S) and Jill (S′).

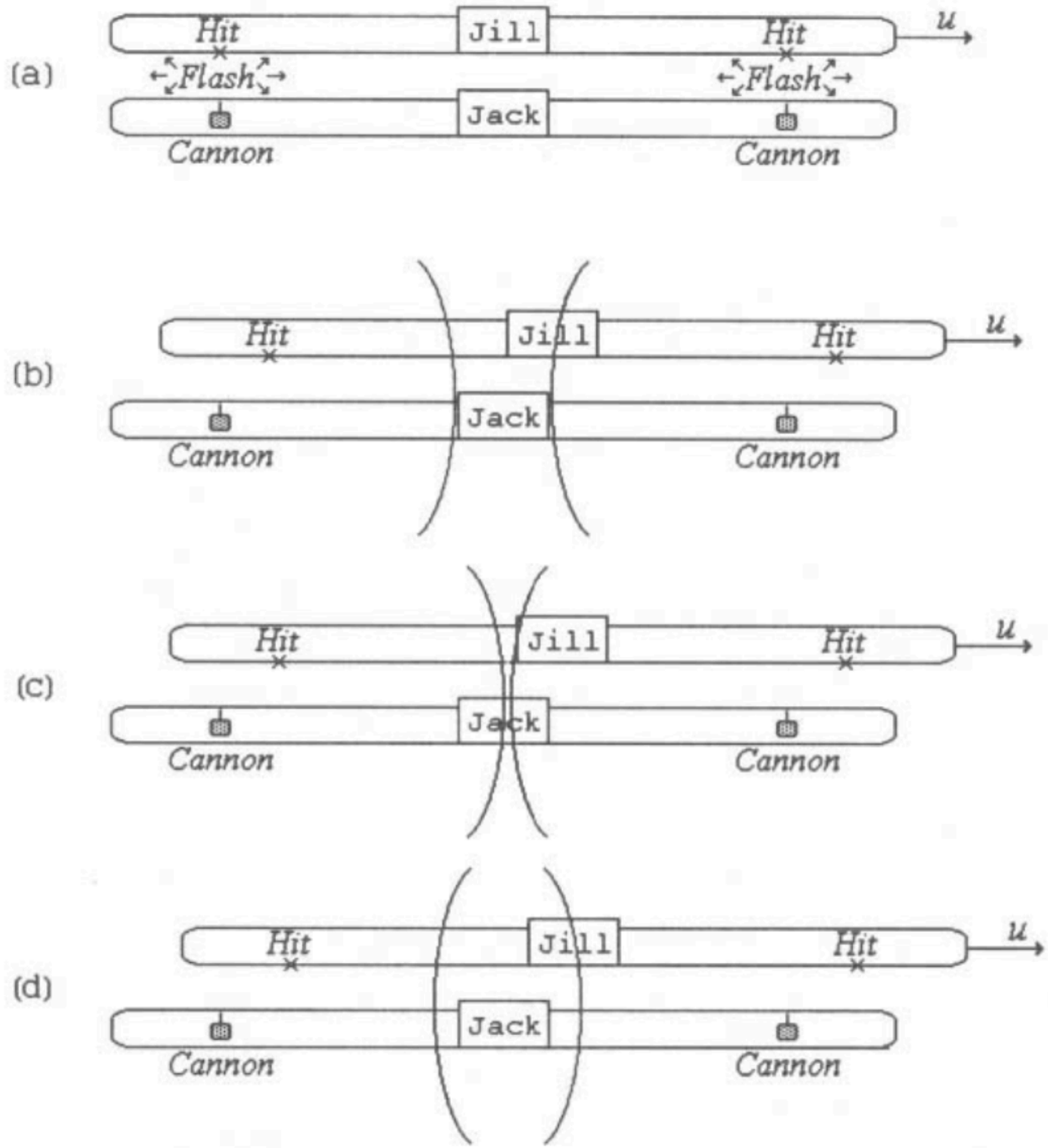


Fig **1.23** The two spaceships from Jack's viewpoint. Jill is moving **to** the right with constant velocity, ***U,*** relative to Jack. (a) The cannons of Jack's ship fire **as** the two ships draw alongside. Wave-fronts from the muzzle flashes radiate outwards in all **directions.(b)**Jill sees the wave-front from the cannon that hit her ship forward of the bridge. (c) Jack **sees** the two flashes from his cannons simultaneously. (d) Jill **sees** the wave-front **from** the cannon that hit her ship behind the bridge.

Jill's ship is moving in a straight line with a constant velocity, u, relative to Jack's.

The spaceships, which belong to rival galactic empires, pass close to one another.

Jack and Jill are the respective commanders of the ships and they each report the incident in their respective logs.

Jack's log: **As the enemy ship drew alongside, two of my cannons fired at point-blank range. I know from the dimensions of my vessel that the cannons are equidistant from the command bridge, one forward and one aft. Since I saw the flashes from the cannons' muzzles at the same moment, I conclude that the cannons fired simultaneously.**

Jill's log: **As we drew alongside the enemy ship, its cannons fired at point-blank range. My ship took two hits, one in front of the command bridge and one astern. Measurements revealed that the hits were equidistant from the bridge.**

Jill sees the flash from the cannon that struck the forward section of her ship before she sees the flash from the cannon-shot that hit the aft section.

Since the light sources are equidistant from her and the speed of the light must be the same in both directions, she must conclude that the hits were not simultaneous. And so she writes:

Jill's log: **I saw the flash from the cannon that hit my ship forward of the bridge before I saw the flash from the one that struck behind the bridge. Since the hits were equidistant from the bridge, I conclude that the enemy's forward cannon fired before his aft cannon.**

Both commanders recorded accurate logs of the incident, each from their own view-point.

But their reports contradict one another; they disagree as to the simultaneity of the firing of the cannons.

According to Jack's inertial reference frame, S, they were fired at the same moment, whereas according to Jill's inertial frame, S′, they were not.

Evidently, Jack and Jill do not share a common time; clocks that are synchronized in Jack's reference frame, do not appear to be synchronized when viewed by Jill from her frame.

In general, inertial observers, whose reference frames are in relative motion, will not agree about the simultaneity of events occurring at different points in space.

The reference frame of each observer appears to the other to be affected by their relative velocity.

If the simultaneity of events occurring in different reference frames cannot be presumed, scientific theories that require or imply instantaneous actions and reactions must be discarded.

Contrary to Newton's Third Law:

> **The forces acting between two charges moving relative to one another are neither equal in magnitude nor opposite in direction.**

In essence, this was Faraday's contention when he insisted that action at a distance must be mediated by a field.

The field compensates for any imbalance in action or reaction that arises out of the fact that changes in one component of a system are not transmitted instantaneously to all the other components.

For example, when considering the conservation of linear momentum in a dynamic system of interacting particles, such as a system of charged particles moving relative to one another, linear momentum must also be ascribed to the field in order that the system's total linear momentum be conserved at all times.

The field acts as a flexible reserve of energy and momentum, available at all times to ensure their overall conservation whatever changes are occurring in the system.

Thus, the relativity of time and simultaneity and the concept of the field can be regarded as corresponding concepts.

To return to the paradox of the observers on the two spaceships S and **S'.**

The astonishment felt at the suggestion that observers at different points in space and who are moving away from each other both see the light from the flare in an identical form, arises from the imaginary reference frame high above the spaceships into which our mind's eye projects itself and from where it views the event.

In this imaginary frame, time is absolute and does indeed "flow equably without relation to anything external"; all observers, whatever they are doing and wherever they are, share the same common time.

Viewed from such a frame, the only way the light from the flare can appear in an identical form to the observers on the two spaceships, S and S′**,** is if they are simultaneously at the same point in space.

But in the physical world we inhabit, and with which science is concerned, such reference frames do not exist. In Einstein's real world**,** simultaneity is relative not absolute and observers cannot be distinguished by means of electromagnetic phenomena.

In the physical reality described by science, it is not the constructs such as time that "flow equably without relation to anything external" but the natural phenomena.

It is the natural phenomena and not our mental constructs that are the invariants of the physical sciences.

1.3.6 The Lorentz Transformation

Some years before the publication of the special theory of relativity, Lorentz had realized intuitively that in order for the laws of electromagnetism (Maxwell's equations) to be covariant and for all inertial observers, stationary or moving, to be electromagnetically equivalent and indistinguishable, the time measure, t', used in the moving system S' must be different from that, t, employed in the stationary system S.

After many laborious calculations, Lorentz succeeded in developing a new transformation which incorporated this idea -The Lorentz Transformation- the equations of which are:

$$x' = \gamma(x - ut) \qquad (1.21) \qquad\qquad x = \gamma(x' + ut') \qquad (1.22)$$

$$y' = y \qquad (1.23) \qquad\qquad y = y' \qquad (1.24)$$

$$z' = z \qquad (1.25) \qquad\qquad z = z' \qquad (1.26)$$

$$t' = \gamma\left(t - {ux}/{c^2}\right) \qquad (1.27) \qquad\qquad t = \gamma\left(t' + {ux'}/{c^2}\right) \qquad (1.28)$$

where the factor γ is given by the equation:

$$\gamma = \sqrt{\frac{1}{1 - {u^2}/{c^2}}} \quad (u < c) \qquad (1.29)$$

In the second part of his 1905 paper, Einstein showed that the Lorentz transformation could be derived directly and quite simply from the two postulates of the special theory of relativity.

This demonstrated the underlying compatibility of Maxwell's electromagnetism and special relativity; they both assume the relativity of time.

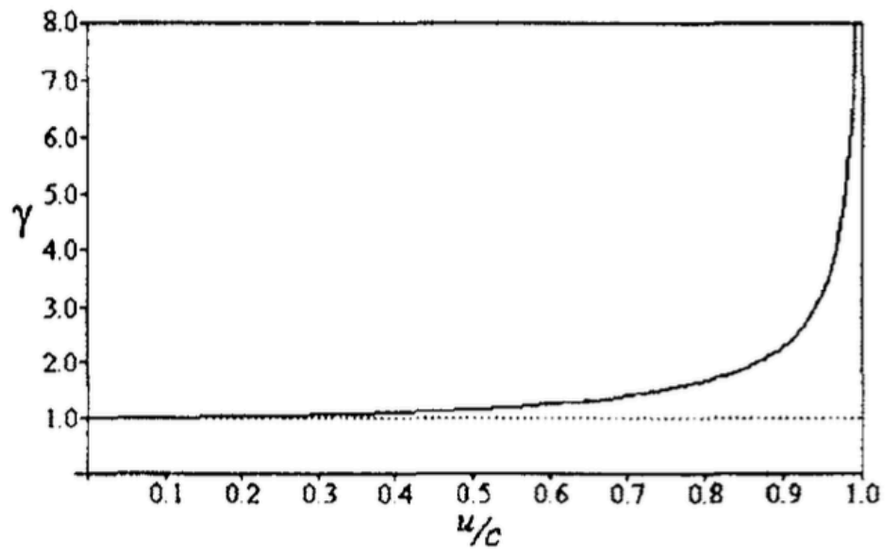
It is Newtonian (classical) mechanics that is the odd man out.

Nevertheless, classical mechanics can still be used to describe everyday motions - in the human sense.

The reason for this is that when the relative velocity, u, between the inertial observers is very much smaller than the speed of light in *vucuo*, c, as it is in everyday human experiences, the equations of the Lorentz transformation reduce very closely to those of the Galilean transformation.

At these low velocities, the factor γ in the Lorentz transformation is practically equal to one (Fig 1.24).

Fig 1.24. Graph of the factor γ against the ratio, u/c, between the relative velocity of the reference frames u and the speed of light in vacuo, c.



In practice, the errors in the predictions of classical mechanics at these velocities are far too small ever to be detected.

Only at relative velocities above 0.5c, half the speed of light (150,000 km/s) do the flaws inherent of classical mechanics begin to be detectable.

Notwithstanding the utility of classical mechanics at low velocities, in order to retain the integrity and unity of physics, a new mechanics that is valid at all velocities and is compatible with special relativity is required.

The laws of this new mechanics - Relativistic Mechanics - are covariant under the Lorentz transformation and are so formulated that the sciences of mechanics and electromagnetism concur.

Digression: The Lorentz Transformation and Special Relativity

As the following discussion shows, the Lorentz transformation for two inertial reference frames, S and S', moving along the XX' axis with a relative velocity of u, can be shown, under certain reasonable premises, to be implicit in the postulates of special relativity.

It is a reasonable assumption that in any transformation the relationship between the coordinates **x** and x' will be of the form:

$$x' = k(x - ut) \tag{1.30}$$

where *k* is a constant of proportionality that is not a function of the space or time coordinates but which may be a function of the relative velocity, u.

The choice of this form arises from a number of considerations:

1. **The relationship between the coordinates x and x' is linear, such that a single event in frame S corresponds to a single event in frame S′;**
2. **The relationship is simple, and simple relations are preferable;**
3. **The relationship can reduce to the corresponding equation of the Galilean transformation, $x' = x - ut$.**

The first postulate of special relativity, the principle of relativity, requires that the laws of physics have the same form in all inertial systems.

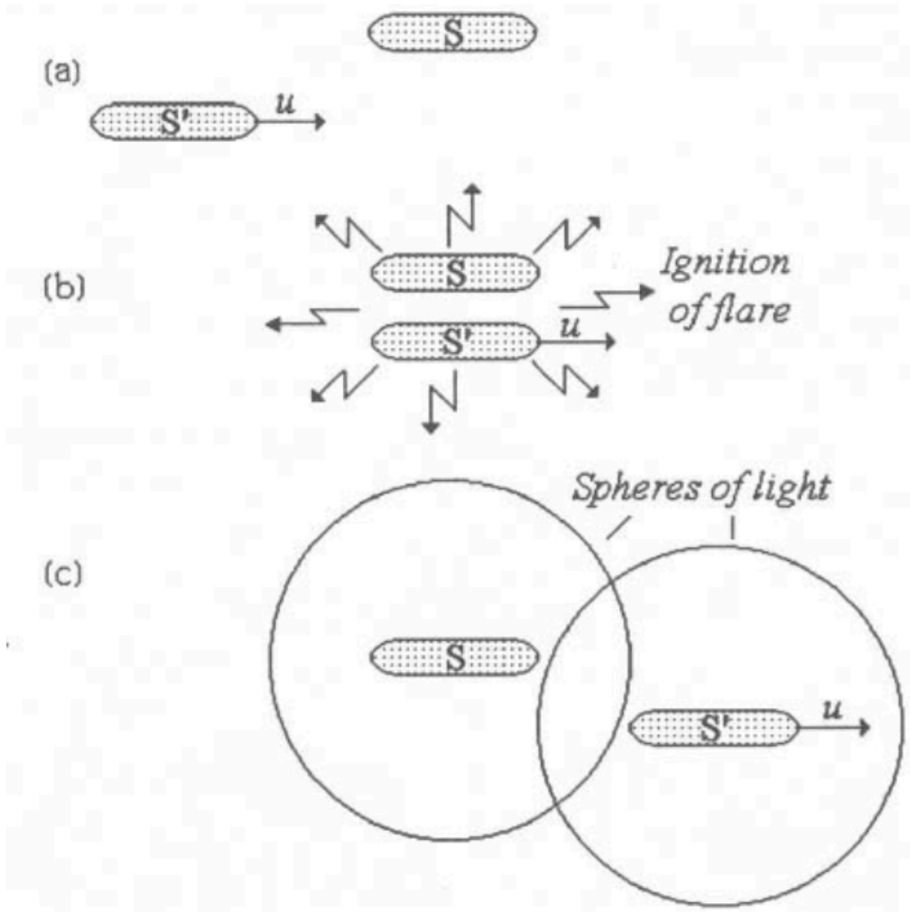
Therefore,. in order to obtain the appropriate relationship which expresses the coordinate x in terms of x' and ***t'***, just the sign of the velocity, u, in the above relationship has to be changed (because of the difference in the direction of their relative motion from the view-point of the other reference frame):

$$x = k(x' + ut') \tag{1.31}$$

There is no relative motion parallel toY or Z and so y=y' and $z = z'$.

The second postulate of special relativity, the constancy of the speed of light, requires that a ray of light observed from each of the reference frames, S and S', must propagate with the same velocity, c.

Suppose, that at the moment the two observers in Fig 1.21 meet and synchronize their clocks, a flare is ignited and the two observers measure the velocity of the propagation of the light from it along the XX' axis.



At a given subsequent moment, the wave-front from the flare will have reached point **x** in the coordinate system of observer S and point x' in that of observer S'.

However, because of the constancy of the speed of light each observer must measure the same value for this velocity. i.e.,

$$c = x/t = x'/t' \tag{1.32}$$

Eliminating **x, x',** t and t' from equations (1.30), (1.31) and (1.32) gives:

$$k = \sqrt{\frac{1}{1 - u^2/c^2}} \tag{1.33}$$

Thus, the factor of proportionality, k, derived from the postulates of special relativity, is identical to the factor γ that appears in the equations of the Lorentz transformation.

The proposed transformation equations, (1.30) and (1.31), are thus identical to the corresponding equations of the Lorentz transformation.

1.3.7 Relativistic Mechanics-Kinematics

The essential feature of the relativistic mechanics is its compatibility with the Lorentz transformation. In the realm of kinematics, we will investigate how this affects the measurement of three physical quantities: distances (spatial displacements) and time intervals; and a new quantity called the *spacetime interval*.

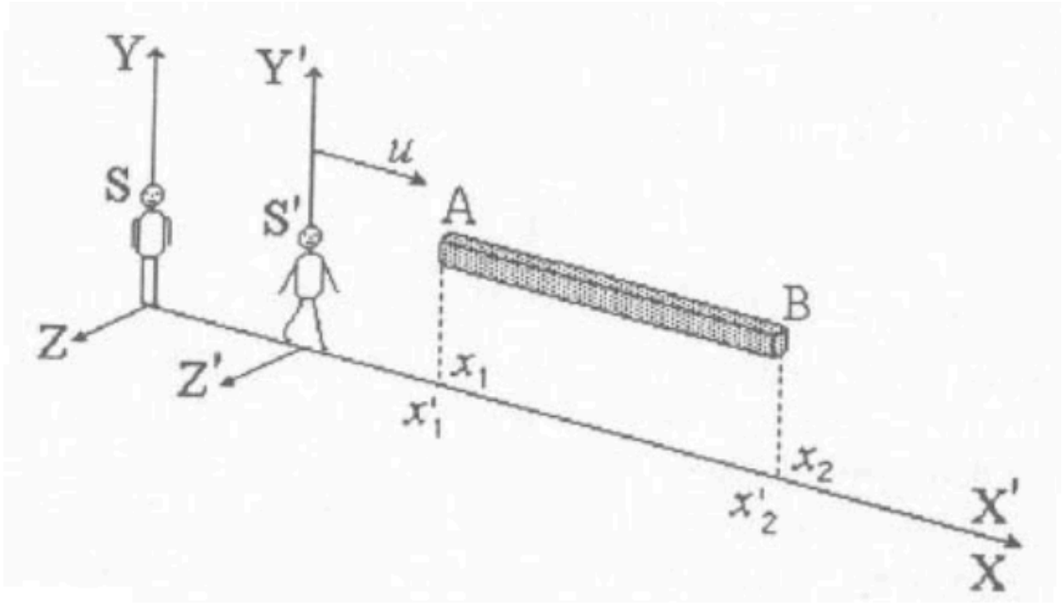
1. <u>The Measurement of Distances - *Length Contraction*</u>

Under the classical Galilean transformation, the measured size of an object is unaffected by the relative motion of inertial observers; distances and displacements are invariants.

However, as we shall see, under the Lorentz transformation, both of these measures are relative quantities.

Consider a rod, AB, that is at rest in the S reference frame (Fig 1.25).

Fig 1.25 Two points, A and B, marking the ends of a rod in the stationary reference fiame ***S***. The **length** of **the rod** according to **the** coordinate system ***S*** is $L_0 = x_2\text{-}x_1$, and according to the coordinate system S′ is $L = x'_2\text{-}x'_1$.



The length of the rod**,** viz**.** the distance between the points A and B parallel to the XX' axis**,** is measured by an observer in the reference frame S and by another in the moving frame S' (Fig 1.25**).**

What is the relationship between the rod's length, L_0**,** measured by the observer in frame S and its length, L, measured by the moving observer from the S' frame?

According to the S coordinate system the length of the rod is:

$$L_0 = x_2 - x_1 \tag{1.34}$$

and according to the S' coordinate system its length is:

$$L = x'_2 - x'_1 \tag{1.35}$$

Substituting the appropriate values for x' and t' given by the Lorentz transformation - equation (1.22) - in place of the x-values in equation (1.34) gives:

$$\begin{aligned} L_0 &= \gamma\left(x_2' + ut'\right) - \gamma\left(x_1' + ut'\right) \\ &= \gamma\left(x_2' - x_1'\right) \end{aligned} \tag{1.36}$$

Combining equations (1.35) and (1.36) gives:

$$L_0 = \gamma L \tag{1.37}$$

The factor γ is always equal to or greater than one. Consequently, in general, $L_0 \geq L$.

The rod has the same dimensions in the Y and Z directions for both observers, there being no relative motion in this direction.

The rod's length, L_0 , measured from the reference frame at rest relative to it, is called its *proper length*.

The length, *L*, measured from the moving reference frame in the direction of its motion, is always shorter than the proper length, L_0, measured from the stationary frame.

Thus, stationary objects appear to be foreshortened when viewed from moving reference frames.

The phenomenon is called the *Lorentz-Fitzgerald Contraction*.

It should be noted that the contraction is a symmetrical phenomenon; if the rod was at rest in the reference frame S', its proper length would be its length as measured in that frame and the observer S would measure the shorter length

2. The Measurement of Time Intervals - *Time Dilation*

The second aspect of relativistic mechanics that we will examine is the effect of their relative motion on the time intervals that inertial observers measure.

Suppose, that observers in the S and S' reference frames possess identical clocks, which they use to measure the time that elapses between two events that occur in the S' frame of reference at a point whose coordinates are (x' , y', z') (Fig 1.26).

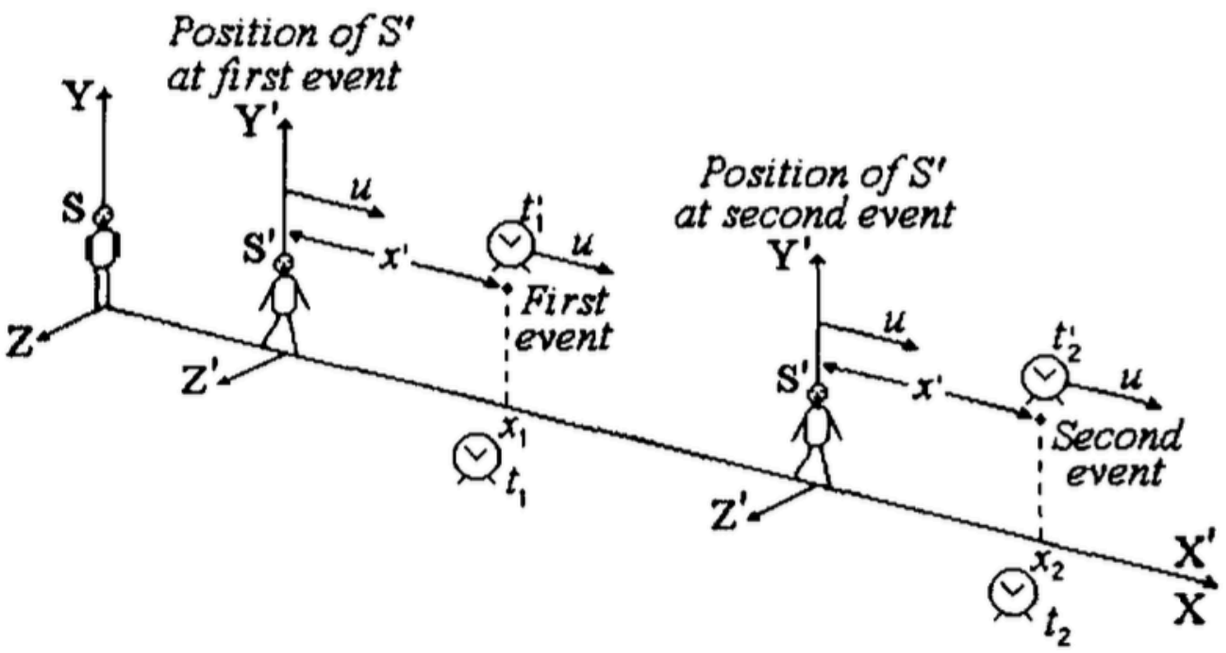


Fig 1.26 Measuring the time interval between two events that occur in the S' frame of reference. Both events occur at the same point whose coordinatesare (x', y', z') in the S' coordinate system but at different points, (x_1, y, z) and (x_2, y, z), in the S system. Observer S needs two synchronized clocks. one at x_1 and the other at x_2. in order to measure the time interval, $\Delta t = t_2 - t_1$, between the vents. However S', who is at rest relative to the events, and sees them both as occurring at the same point, can measure the time interval, $\Delta t_0 = t'_2 - t'_1$, between them with a single clock. S' measures the proper time interval.

In general, an observer at rest relative to a sequence of events can measure the time interval between them with a single stationary clock.

This time interval, Δt_0 , is called the *proper time interval*.

Relative to an observer in the S' reference frame, the two events occur at the same point and so in order to measure the time interval between them he requires just one clock, placed at the point (x', *y*', *z*') in his coordinate system.

Thus the time interval between the two events measured by the observer in the S' reference frame is the proper time interval.

The situation of the observer in the other reference frame is different.

Because the S' reference frame is moving relative to him, the observer in the S reference frame finds that the events occur at two different points in his coordinate system, designated, respectively, by the coordinates (x_1, *y, z)* and (x_2, *y*, z).

Consequently, the observer in the S system requires two synchronized clocks, one at each point, in order to measure the time interval, Δt, between the events.

According to the observer in the S coordinate system, the time that elapses between the events is:

$$\Delta t = t_2 - t_1 \tag{1.40}$$

and according to the observer in the S′ system it is:

$$\Delta t_0 = t_2' - t_1' \tag{1.41}$$

Substituting equation (1.28) of the Lorentz transformation in equation (1.40) gives:

$$\begin{aligned} \Delta t &= \gamma\left(t_2' + {ux'}/{c^2}\right) - \gamma\left(t_1' + {ux'}/{c^2}\right) \\ &= \gamma\left(t_2' - t_1'\right) \end{aligned} \tag{1.42}$$

Hence

$$\Delta t = \gamma \Delta t_0 \tag{1.43}$$

However, the factor γ is always equal to or greater than one; $\gamma \geq 1$.

Consequently, in general, $\Delta t \geq \Delta t_0$.

Measured from a stationary reference frame such as S, the time, Δt, that elapses between events occurring at the same place in a moving reference frame, S', is always longer than the proper time interval, Δt_0, measured by an observer in the S' frame.

The phenomenon is called time dilution and is exhibited by all physical processes including chemical reactions and biological growth.

The phenomenon of time dilation has been confirmed in numerous experiments.

For example, in investigations of the average life-times of the unstable particles called pions, it is found that at rest, pions decay on average after 26 nano-seconds; this is their proper average life-time.

However, their measured average life-time when moving at a speed of 0.99c relative to the laboratory is, according to the laboratory's clocks, 7.1 times longer.

Thus, whereas an observer riding one of the pions has to wait, on average, just 26ns for it to decay, an observer in the laboratory has to wait 185ns.

To put this another way, relative to the stationary clocks in the laboratory, the moving pion's own clock is slow; it shows the passage of just 26ns when, according to the laboratory clocks, 185ns have already passed.

This result is often expressed by the adage "moving clocks run slowly".

In general, the shortest time interval observed between events is the proper time interval measured from the reference frame in which the events are occurring, i.e., by a clock at rest relative to the events.

The time intervals measured by clocks in reference frames moving relative to the events are always longer.

An important corollary of time dilation is that the proper frequency, f_0, of periodic events, i.e., their frequency as measured from the reference frame in which they occur, is always higher than the frequency measured from a frame moving relative to the events.

For example, consider measurements of a patient's heart-beat made by two doctors, one who is at rest relative to the patient and one hurrying past.

Whereas the stationary doctor measures the patient's proper heart-beat, the running doctor will measure a marginalIy lower heart-beat.

If the process of ageing depends, among other things, on the rate of one's heart-beat, the patient is ageing faster relative to the stationary doctor than he is relative to the running doctor.

As long as keeps moving, the running doctor is not aware of this and it will not affect his diagnosis.

Only when he returns and finds the patient to be older than he expected will he become aware of the difference between his measurement and that of his stationary colleague.

Though time is in fact a relative measure - a construct and not an absolute - not all our intuitive ideas about it are flawed.

Although the laws of physics do not distinguish between forward and backwards directions of time, time does not flow backwards for any observer.

Similarly, events occur in the same order for all observers, though each may measure a different time interval between them.

By the same token, a distant observer, whatever his motion, cannot observe an event before an observer who is closer to it.

And, finally, we cannot look into the future.

3. The Fourth Dimension

Eliminating the relative velocity, u, from the equations of the Lorentz transformation by simple algebra, gives the following relationship between the space and time coordinates of the two observers S and S':

$$(ct)^2 - \left(x^2 + y^2 + z^2\right) = (ct')^2 - \left(x'^2 + y'^2 + z'^2\right) \tag{1.44}$$

The new physical quantity, I , defined by the equation

$$I = (ct)^2 - \left(x^2 + y^2 + z^2\right) \tag{1.45}$$

is an invariant quantity called the *interval*.

Its significance in relativistic mechanics is analogous to that of distance in classical mechanics; it is the quantity that separates events in the Einsteinian conception of motion.

On their own, neither space intervals (spatial distances) nor time intervals are invariant.

However, the quantity I, which is compounded of both time and space, and in which the time coordinate t has the same standing as the three space coordinates x, y and z, is invariant.

The notion underlying this concept is that objects should be regarded as occupying points in a four-dimensional geometrical entity called *spacetime*.

Though, in classical physics, space and time coordinates can be considered separately, in relativistic physics they are treated together under the single heading of spacetime.

Hence, whereas classical mechanics employs *three-vectors* , i.e., vectors having three components, in its formulations, relativistic mechanics requires the use of *four-vectors*, i.e., vectors possessing four components, one for each of the possible directions, x, y , z and t .

Events occurring in spacetime, as well as motion through it, can be depicted in a graphical description, a spacetime diagram**,** first suggested by Minkowski. Fig 1.27 is the Minkowski spacetime diagram of the stationary observer S; only motions parallel to the X axis are shown.

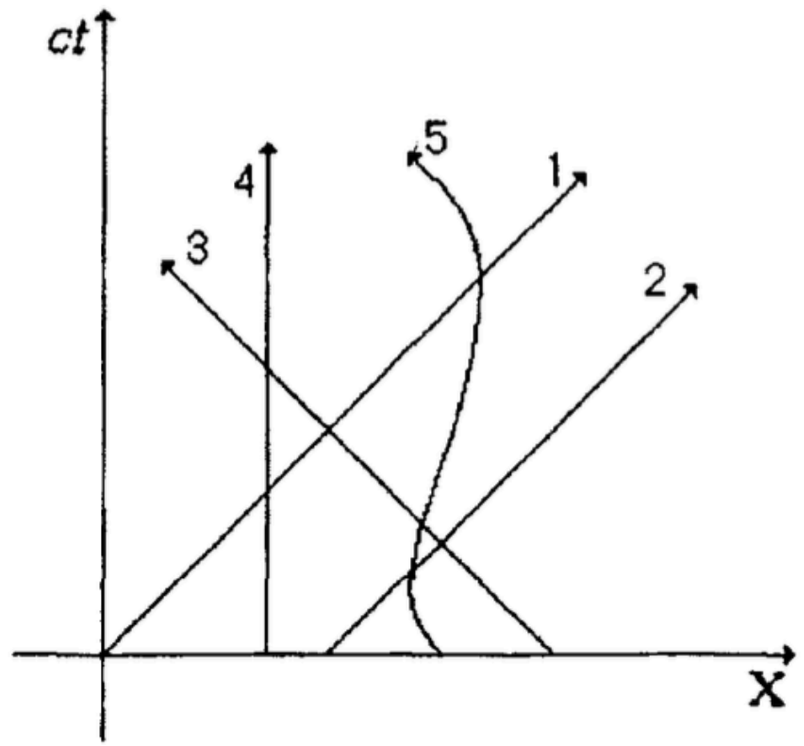


Fig 1.27 Minkowski diagram for the observer S. The lines numbered 1, 2 and 3 are the world-lines of light signals propagated from different points in different directions. Note that the spacetime interval, $(c\Delta t)^2 - (\Delta x)^2$, between any two points on these world-lines is zero. Line 4 represents a body at rest and line 5 one whose motion is changing with time, i.e., accelerated motion.

The horizontal axis in the diagram is the space axis and corresponds to the X axis of the observer S**.**

The vertical axis is the time axis; it is scaled in units of ct - units of length - so that the two axes should have the same units.

A line or curve drawn on the diagram represents the motion of a particle or signal through spacetime; such lines or curves are called *world-lines***.**

The velocity, u, of a particle or a signal is inversely proportional to the slope of its world-line in the observer's spacetime diagram:

$$u = c \cdot \frac{\Delta x}{\Delta(ct)} \tag{1.46}$$

Hence, a light ray, i.e., a signal moving with velocity c, is represented by a line of slope 45°: the lines numbered 1, 2 and 3 in Fig 1.27.

A body at rest, u = 0, is represented by a straight vertical line: line number 4.

A body whose motion undergoes change is represented by a curved or broken line: line number 5.

Minkowski diagrams can also be used to elucidate the nature of physical causality, i.e., what can and cannot be the cause or the result of a physical event.

Clearly, one event cannot be the cause of another before information of its occurrence arrives at the point where the second event is to occur.

The maximum speed at which information can be transmitted, c, is represented in a spacetime diagram by the surface of a cone of slope 45°; the cone is called the *light cone*.

The light cone of an event is divided into the past light cone and the future light cone (Fig 1.28).

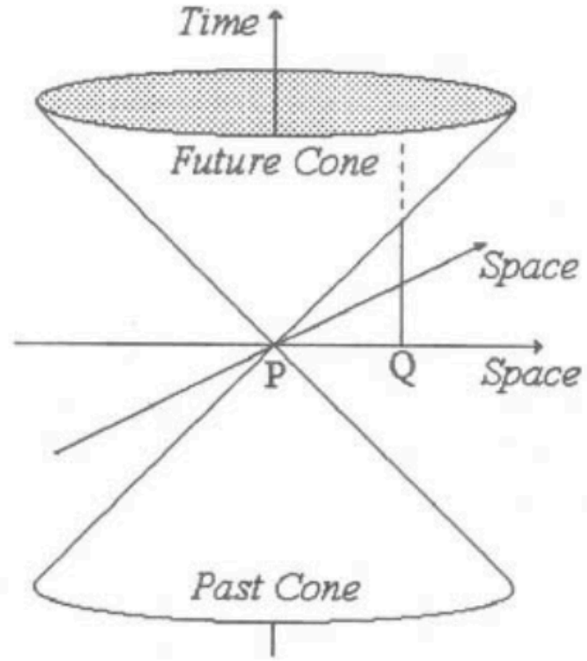


Fig 1.28 The light cone of an event, P. The P event *can* be the cause of any event in its future cone or the result of any event in its past cone It has no effect on, nor is it affected by, events outside its light cone. A stationary body at Q only learns of the P event when its vertical world-line enters P's future light cone.

The P event, represented in Fig 1.28, can affect all the events inside its future light cone - its absolute future - and can be affected by all those in its past iight cone - its absolute past.

However, it cannot affect events outside its future light cone nor is it affected by those outside its past light cone.

A stationary body at Q will not be affected by the P event until, after the passage of some time, its vertical world-line enters P's future light cone.

Likewise, for an event occurring at Q to be the cause of the P event would require an instantaneous interaction, which is an impossibility.

The Minkowski Diagrams of Different Observers - The Twins Paradox

We will start by explaining the construction of the Minkowski diagrams of two inertial observers and then use them to elucidate what has become known as the *Twins Paradox*.

1. <u>The Minkowski Diagrams of the Two inertial Observers S and S'.</u>

Since S' moves with constant velocity, u , parallel to the X axis, the slope of his world-line in the Minkowski diagram of the observer S is c/u.

However, at every point on this line $x' = 0$.

Hence, it defines the ct' axis of the Minkowski diagram of the moving observer, S' (Fig1.29).

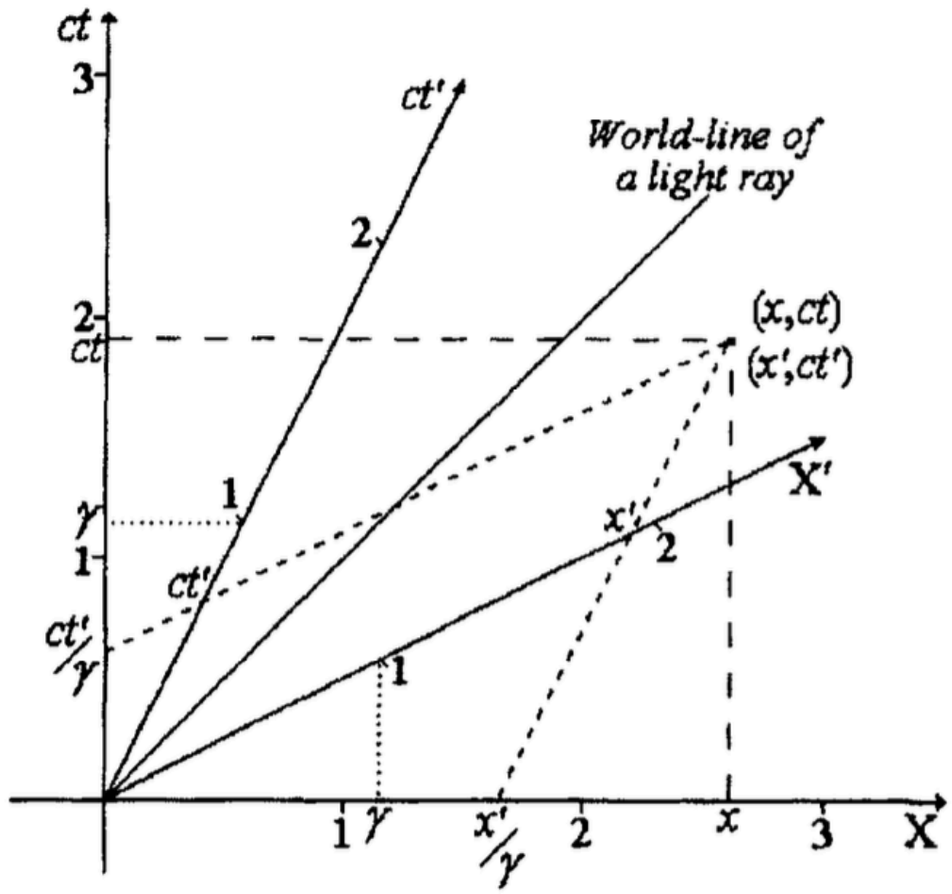


Fig 1.29 Minkowski diagrams of the two observers, S and S'. In the orthogonal system of the observer S, the gradient of the X' axis is u/c and that of the ct' axis is c/u . A point with the spacetime coordinates (x, ct) according to the observer S has the coordinates (x', ct') according to the observer S'.

The speed of light in *vacuu* has the same value for all inertial observers.

Thus, a ray of light will be represented by the same world-line in the Minkowski diagrams of both *S* and S' .

In the diagram of the observer S, the world-line of a light ray bisects the angle between the axes; it is reasonable to assume that the same will be true for the diagram of the observer S'.

This defines the orientation of the X' axis.

Note, that a consequence of this is that the angle between the X' and ct' axes is not a right-angle.

Because of their relative motion, the scaling of the axes of the two observers differs.

We employ the Lorentz transformation in order to scale the X' and ct' axes,.

From equation (1.22), the x coordinate that corresponds to the point whose coordinates are x' = 1 and t' = 0 in the diagram of the observer S' is

$$x = \gamma(x' + ct') = \gamma(1+0) = \gamma$$

i.e.. the point x' = 1 lies on the X' axis immediately above the point x = γ on the X axis.

SimilarIy, from equation (1.28), the corresponding ct coordinate of the point having coordinates x' = 0 and ct' = 1 is

$$ct = c\gamma\left(t' + {ux'}/{c^2}\right) = c\gamma\left({1}/{c} + 0\right) = \gamma$$

i.e., the point ct' = 1 lies on the ct' axis directly opposite the point $ct = \gamma$ on the ct axis.

In this way, the spacing of the integers on the axes of the S' diagram is defined.

To find a point's coordinates (x', ct') relative to the moving observer S', lines (the broken lines in Fig 1.29) are drawn from the point parallel to the X' and ct' axes.

The coordinates x' and ct' are respectively defined by the points where these lines cut the X' and ct' axes.

The extension of these two lines crosses the X and ct axes at the points x'/γ and ct'/γ , respectively.

2. The Twins Paradox

What has become known as the *Twins Paradox* was first propounded by Einstein in 1911:

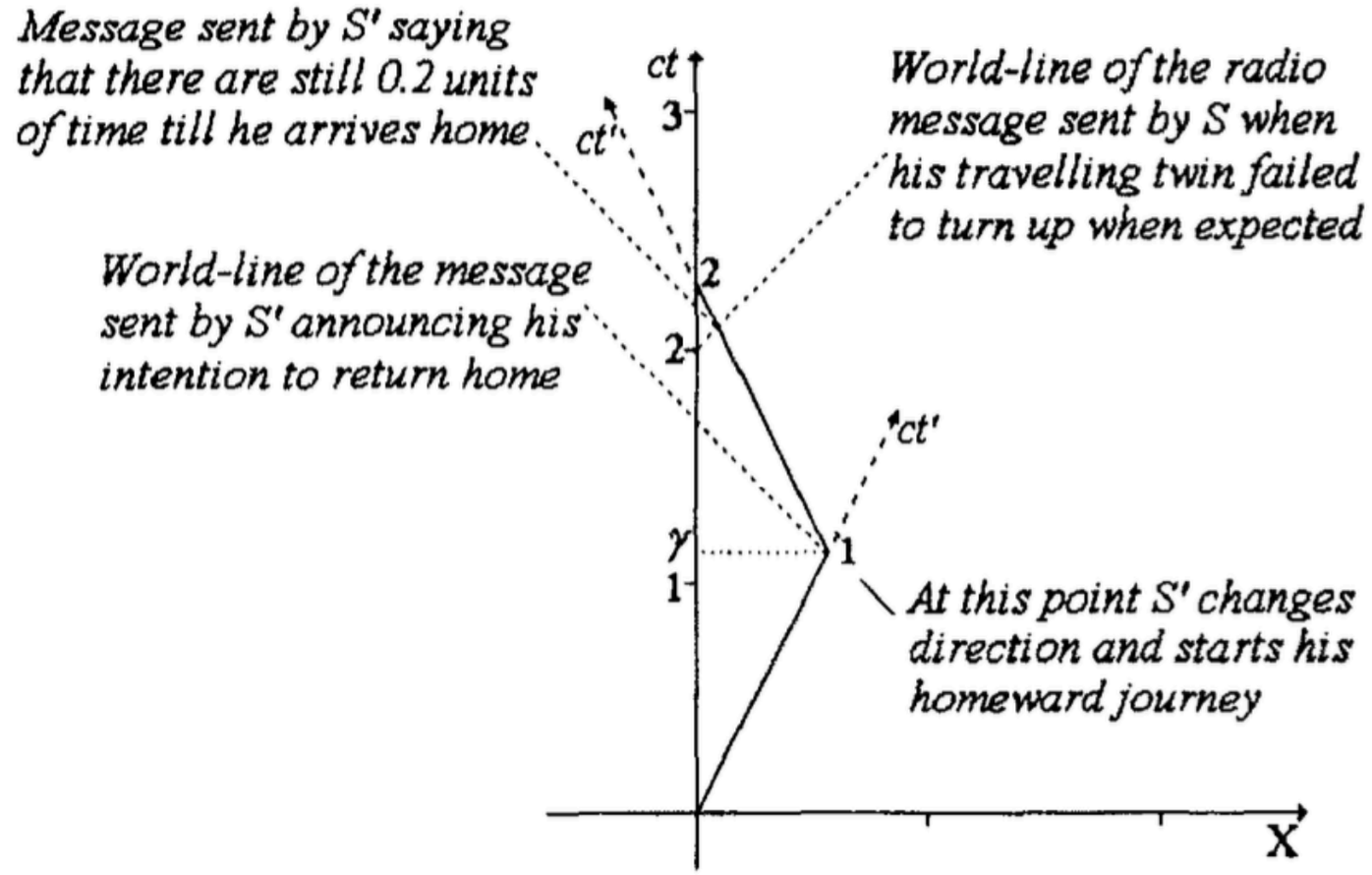
> ***If we placed a living organism in a box [a spaceship] ... one could arrange that the organism, after any arbitrary lengthy flight, could be returned to its original spot in a scarcely altered condition, while corresponding organisms which had remained in their original positions had long since given way to new generations.***

If the traveling organism is a man and the one left behind on earth is his twin, the traveling twin finds, upon his return home, that his brother/sister has aged more than himself.

It might appear at first that the phenomenon can be simply explained by time dilation; the traveling twin's clock-the rate at which his body ages-runs more slowly than that of the twin left on earth and so he has aged less that his stationary twin.

Let the traveling twin's velocity u = 0.5c. Each twin's world-line defines the time axis in his Minkowski diagram; that of the stationary twin, S, is the ct axis, that of the traveling twin, S', is the ct' axis. However, as the diagram in Fig 1.30 shows, their respective world-lines are not interchangeable.

Fig 1.30 The respective world-lines, ct and ct', of the homebound twin, S, and traveling twin, S', and of their radio messages. The traveling twin moves at a speed of 0.5c relative to the stationary twin. The slope of the world-lines of each radio message is 45°. After traveling for a single unit of time as measured by his clock, the traveling twin changes course and makes for home. He arrives home after t' = 2 units of time according to his clock but after t = 2.3 units of time according to the stationary twin's clock



The orientation of the traveling twin's time axis must change somewhere in spacetime if he is to return home; this change is equivalent to an acceleration. Whereas the homebound twin remains in an inertial frame throughout, in order to return home the traveling twin must undergo an acceleration. Thus, their situations are not symmetrical.

Suppose, that at the moment he changes course and makes for home-1 time unit into his journey according to his clock-the traveling twin, S', transmits a light signal (a radio message) announcing his intention of returning and that he estimates he will be arriving home after having been away for 2 units of time. This signal is received by the stationary twin, S, 1.73 units of time, according to his clock, after the traveling twin's departure.

However, motion is relative; either twin can contend that the other was the traveller and so each can contend that the other should be younger, which is obviously impossible.

This is the essence of the paradox.

The twins' situations appear to be symmetrical and interchangeable. Are they so in fact?

When 2 units of time have elapsed, according to the stationary twin's clock, the traveller has still not yet returned.

Worried as to his peripatetic twin's fate, the stationary twin sends him another message.

This message is received by the traveling twin at a moment, according to his clock, that is only about 1.8 units of time after his departure.

So he sends back another message asking why all the panic: "There are still 0.2 units of time left before I expect to arrive home."

The message is received just before S′ gets back.

When the twins meet, it becomes apparent that whereas the traveling twin’s clock has recorded the passage of 2 units of time since he left home, the stationary twin’s identical clock-his genetically identical body-has recorded the passage of 2.3 units of time; S has aged more than S′.

Each twin’s proper time, the time according to which he himself ages, is measured by his individual time axis.

1.3.8 Relativistic Mechanics - Dynamics

The impact of the Einstein's revolutionary conception of motion on the study of dynamics was no less profound than it was on the science of kinematics.

Like the relativistic kinematics, the new *relativistic dynamics* has stood every experimental test to date.

Perhaps its most celebrated consequence was the notion of the underlying equivalence of mass and energy expressed by the equation $E = mc^2$**.**

In the following discussion, we will first examine how the consequences of this new conception of motion on our understanding of four familiar physical quantities - linear momentum, mass, force and kinetic energy - and how the notion of mass-energy equivalence arises from it

1. <u>Linear momentum</u>

In classical mechanics, the linear momentum, **p,** of a particle of mass m moving with velocity **u** relative to a given reference frame is defined as:

$$\mathbf{p} = m\mathbf{u} \tag{1.47}$$

In experiments conducted at low velocities, linear momentum, thus defined, is generally found to be conserved.

However, in experiments with high energy particles moving at speeds approaching the speed of light, it is found that linear momentum, as defined classically, is not conserved.

In explaining why this is so, we will restrict ourselves to uniform motion parallel to the X axis.

In the classical conception, a particle's velocity, u, is defined as the ratio between its displacement, Δx, relative to an inertial observer, and the period of time measured by the same observer, Δt, during which this displacement occurs:

$$u = \frac{\Delta x}{\Delta t} \tag{1.48}$$

This gives for the linear momentum of the particle:

$$p_{classical} = m\frac{\Delta x}{\Delta t} \tag{1.49}$$

However, this definition of linear momentum ignores the relativity of time.

It assigns an unjustified pre-eminence to the time interval, Δt, measured by the observer.

As opposed to this, the relativistic definition of linear momentum expresses the particle's linear momentum, $p_{relativistic}$, in terms of the proper time interval, Δt_0, measured in the reference frame of the particle itself:

$$p_{relativistic} = m\frac{\Delta x}{\Delta t_0} \tag{1.50}$$

Replacing the proper time interval, $\Delta t_0 = \Delta t / \gamma$, in this equation by the observer's time interval gives

$$p_{relativistic} = m\gamma\frac{\Delta x}{\Delta t} \tag{1.51}$$

Combining equations(1.49) and(1.51) produces the relationship:

$$p_{relativistic} = \gamma p_{classical} \tag{1.52}$$

Thus, in relativistic mechanics,the definition of linear momentum is:

$$p_{relativistic} = \gamma mu \tag{1.53}$$

When inertial observers employ this definition of linear momentum, they find that linear momentum is conserved whatever the speed of the particles.

2. Mass

The relativistic definition of linear momentum can also be written in the form:

$$p_{relativistic} = m_{relativistic} u \tag{1.54}$$

where

$$m_{relativistic} = \gamma m_{classical} \tag{1.55}$$

Since $\gamma \geq 1$, $m_{relativistic} \geq m_{classical}$.

* A particle's *classical mass*, $m_{classical}$, is its mass as measured from its own reference frame, i.e., from a reference frame at rest relative to the particle. It is the mass assigned to it by classical mechanics. In relativistic mechanics, it is usually designated m_0 and termed the *rest mass* or *proper mass* of the particle.
* A particle's *relativistic mass*, $m_{relativistic}$, is its mass measured from a reference frame moving at a velocity u relative to the particle; it is usually designated simply m in relativistic mechanics.

A particle's relativistic mass depends on the factor γ, which is itself a function of the observer's motion.

Thus, in relativistic mechanics, mass is not an invariant as it is in classical mechanics; it is a relative quantity whose value depends on the relative motion of the observer and the particle.

At speeds below 0.5c, the value of γ is so close to unity that the increase in mass is usually undetectable and can be ignored.

However, as an object's speed approaches the speed of light, its mass rises asymptotically to infinity and classical mechanics is no longer competent to describe its behavior.

The dependence of an object's mass on its speed relative to the observer has been confirmed experimentally many times.

In a series of experiments carried out by Bucherer in 1909, the ratio, q/m between the charge and the mass of the electron was found to depend on the particle's velocity; lower values were found at high velocities than at low-velocities (Table 1.1).

This was attributed to the relativistic increase in the mass of the particles that resulted from their motion.

Table 1.1: The Dependence of the Charge to Mass Ratio of Electrons on their Velocity

u	γ	Observed charge to mass ratio (10^{11}C/kg)	Proper charge to rest mass ratio (10^{11}C/kg)[45]
0.321c	1.056	1.66	1.75
0.379c	1.081	1.63	1.76
0.430c	1.108	1.59	1.76
0.521c	1.171	1.51	1.77
0.690c	1.381	1.28	1.77

The variation in the charge to mass ratios could have been attributed to decreases in the electric charge carried by the particles and not to increases in their mass.

However, all the evidence suggests that electric charge remains an invariant quantity in relativistic physics just as it was in classical physics.

3. Force

Force is defined in relativistic mechanics, as it is in classical mechanics, as the rate of change of linear momentum:

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} = \frac{d(m\mathbf{u})}{dt} = \frac{d(\gamma m_0 \mathbf{u})}{dt} \tag{1.56}$$

Both γ and u are functions of t.

Consequently, differentiation gives:

$$\mathbf{F} = \gamma m_0 \frac{d\mathbf{u}}{dt} + m_0 \mathbf{u} \frac{d\gamma}{dt} \tag{1.57}$$

This relativistic equation reduces to Newton's Second Law of Motion, $\mathbf{F} = m\mathbf{a}$ in the special case where $\gamma = 1$.

Since, in general, the force has a component $m_0\mathbf{u}\frac{d\gamma}{dt}$, in the direction of the velocity **u**, in relativistic mechanics, a particle's acceleration, $\mathbf{a} = d\mathbf{u}/dt$, is not necessarily parallel to the force acting upon it as it is in classical mechanics.

If the particle moves at a constant speed u, as in uniform circular motion, $\frac{d\gamma}{dt} = 0$. In this case equation (1.57) reduces to:

$$F_{Centripetal} = \gamma m_0 a_N = \gamma m_0 u^2 / R \tag{1.58}$$

where R is the radius of the circle.

It can be shown that if F_x, F_y, and F_z, are the components of a force acting on a particle at rest relative to the inertial observer S, then the components of the force acting on the particle relative to the moving inertial observer S′ are:

$$F'_{x'} = F_x \qquad F'_{y'} = {}^{F_y}\!/_{\gamma} \qquad F'_{z'} = {}^{F_z}\!/_{\gamma} \tag{1.59}$$

4. Kinetic energy

The kinetic energy of a charged particle such as an electron can be expressed in terms of mechanical or electrical measures, depending on the circumstances:

* In terms of mechanical measures such as the particle's mass and velocity. In classical mechanics, this gives the familiar expression $K = {}^{m_e u^2}\!/_{2}$ for the kinetic energy of an electron moving with velocity u in which m_e designates the electron's classical mass.
* In terms of electrical measures such as the particle's charge and the potential difference through which it is accelerated. This gives the equally familiar expression $K = eV$ for the kinetic energy of a particle carrying a charge e that is accelerated through a potential difference V.

Though each definition employs different measures, to be consistent, both should give the same numerical value for the particle's kinetic energy. Equating the two definitions, $K = m_e u^2/2$ and $K = eV$, gives the following expression for the square, u^2 , of the velocity of an electron accelerated through a potential V.

$$u^2 = 2eV/m_e \tag{1.60}$$

If the two definitions are indeed consistent, measured velocities should correspond with those predicted by this equation.

This is borne out at potential differences below ~100,000V.

However, at higher potentials (higher energies), the measured velocities are found to be lower than those it predicts.

Instead of increasing linearly with the potential, *V,* as it should according to equation (1.60), the square, u^2, of their velocity rises asymptotically to c^2 (Fig 1.31).

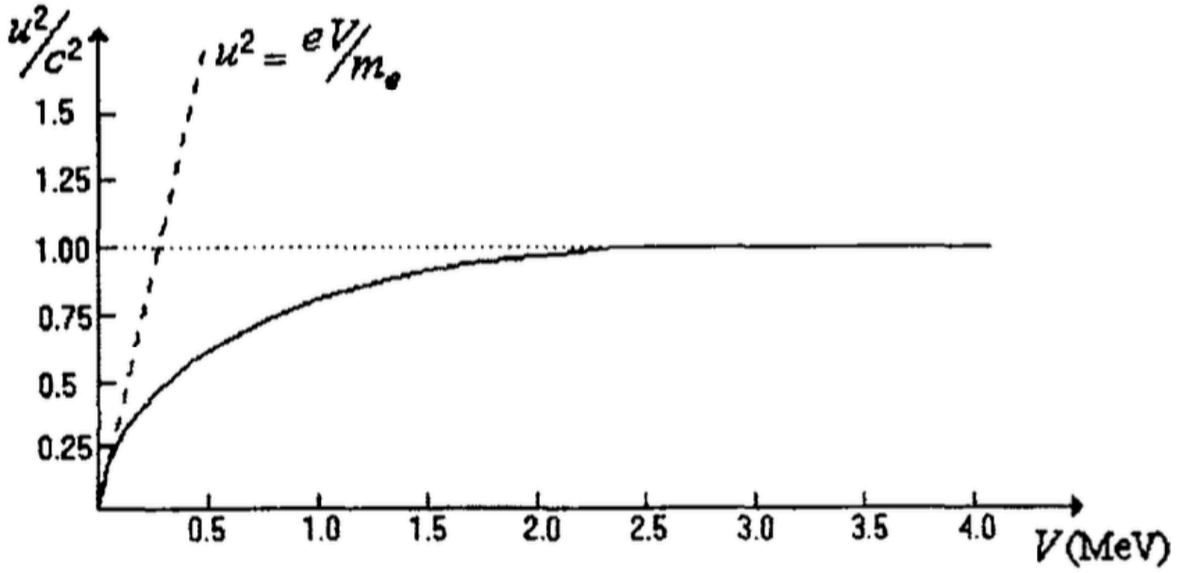


Fig 1.31 The relationship between the square, u^2, of the velocity of an electron - expressed as a fraction of the square of the **speed** of light in *vacuo* - and the potential *V* through which it was accelerated. The broken line shows how the square of the particle's velocity would rise were the classical definition valid at all energies.

The source of the problem is that the classical expression for kinetic energy, $K = m_e u^2/2$, recognizes neither an upper limit on the speed to which the particle can be accelerated nor that its mass increases with its speed.

Whereas these considerations can be ignored for all practical purposes at low energies, they cannot be disregarded at high energies.

At such energies, relativistic effects must be taken into account and a different expression must be employed for the particle's kinetic energy, one that takes these effects into consideration.

Kinetic energy is defined in relativistic mechanics as it is in classical mechanics, namely, as the work done by an external force in increasing the speed of the particle.

Restricting ourselves to forces and motion parallel to the X-axis, this gives the integral:

$$K = \int_0^x F_x dx \tag{1.61}$$

where. F_x is the external force acting on the particle.

When carrying out this integration in relativistic mechanics, account must be taken of the change in the particle's mass that occurs as its speed increases.

This is done by substituting the relevant relativistic definition of force, $F_x = \frac{dp_x}{dt} = \frac{d(mu)}{dt}$ given by equation (1.56), in the integral.

$$K = \int_0^x \frac{d(mu)}{dt}\,dx \tag{1.62}$$

Performing the integration gives the following expression for a particle's kinetic energy,

$$K = (m - m_0)c^2 \tag{1.63}$$

where m is the particle's relativistic mass and m_0 its rest-mass.

Particle velocities forecast by equating this relativistic definition of kinetic energy with the electromagnetic definition are confirmed by experiment at all energies.

At low energies, i.e., when u << c, the relativistic definition of kinetic energy, $K = (m - m_0)c^2$ reduces to the classical definition, $K = m_0 u^2/2$

5. Mass-energy

A corollary of the relativistic definition of kinetic energy is that any change, ΔK, in the kinetic energy of a particle will be accompanied by a corresponding change, Δm, in its relativistic mass, such that:

$$\Delta K = \Delta m c^2 \tag{1.64}$$

The additional mass acquired by the particle by virtue of its increased kinetic energy is qualitatively identical to the particle's rest mass.

Consequently, the latter can be similarly expressed in terms of an intrinsic energy, E_0, called the particle's *rest-energy:*

$$E_0 = m_0 c^2 \tag{1.65}$$

The total energy, E, of the moving particle is the sum of its own intrinsic energy and its acquired kinetic energy

$$E = E_0 + K \tag{1.66}$$

In terms of its rest mass, m_0, and its linear momentum, p , the total energy of the moving particle is also given by the important and useful formula

$$E^2 = \left(m_0 c^2\right)^2 + p^2 c^2 \tag{1.67}$$

In extending this idea, Einstein proposed that a particle's mass is affected not only by changes in its kinetic energy but by changes in any of the forms of energy it may possess, chemical, potential, electrical etc.

On this view, mass is equivalent to energy in whatever form it takes, such that the general relationship between the total energy, E, of a particle and its total relativistic mass, m**,** is given by the equation

$$E = mc^2 \tag{1.68}$$

It follows that the emission of energy from any system, in any form whatsoever, is done at the expense of the system's overall mass, and that the absorption of energy in any form increases the system's overall mass.

Consequently, the separate classical laws of mass and energy conservation are flawed and must be abandoned.

Neither mass nor energy alone are conserved in natural processes.

The quantity actually conserved - mass-energy - combines mass and energy, taking their equivalence into account.

In SI units, the constant of proportionality, c^2, in the equation $E = mc^2$ is approximately equal to 10^{17}.

Accordingly, one kilogram of any substance -iron, cream cheese, sea water, etc. - is intrinsically equivalent to about 10^{17} joules of energy, and vice versa.

Thus, the annihilation of even a minute amount of matter releases considerable amounts of energy: the substantial amounts of energy released by the disintegration of radioactive materials are produced by just minute reductions in their masses.

By the same reasoning, an enormous quantity of energy would have to be absorbed or emitted by a system for any detectable change in its mass to occur: the amounts of energy involved in ordinary chemical reactions are far too small to produce any significant or even measurable change in the mass of the reactants.

The tremendous quantities of energy released in stellar processes involve the destruction of matter on a cosmic scale; it is estimated that to maintain the rate at which it emits energy, the sun loses about 5 million tons of its matter each second (out of its total mass of some 10^{27} tons).

The Transformation of Momentum and Energy

A particle's linear momentum, **p**, and its total energy, E, are relative quantities, measurements of which depend on the inertial reference frame from which they are made.

In classical mechanics the different measurements are reconciled by means of the Galilean transformation. In relativistic mechanics the Lorentz transformation must be employed.

What form do the equations of the Lorentz transformation take in the case of linear momentum and total energy?

Linear momentum is a vector quantity; the components of a particle's linear momentum relative to the inertial observer S are p_x, p_y, and p_z. The magnitude, p, of the particle's momentum is given by

$$p^2 = p_x^2 + p_y^2 + p_z^2 \tag{1.69}$$

Substitution in the relativistic expression for the particle's total energy relative to the same observer, $E = (m_0^2)^2 + p^2c^2$, followed by rearrangement of the equation gives

$$E^2/c^2 - \left(p_x^2 + p_y^2 + p_z^2\right) = m_0^2 c^2 \tag{1.70}$$

Both m_0, the particle rest mass, and c, the speed of light in *vacuo*, are invariants. Hence, the corresponding expression relative to the inertial observer S′ must be

$$E'^2/c^2 - \left(p'^2_{x'} + p'^2_{y'} + p'^2_{z'}\right) = m_0^2 c^2 \tag{1.71}$$

where p'_x, p'_y and p'_z are the components of the particle's linear momentum relative to S′ and E′ is its total energy. Equating equations (1.70) and (1.71) gives

$$E'^2/c^2 - \left(p'^2_{x'} + p'^2_{y'} + p'^2_{z'}\right) = \frac{E^2}{c^2} - \left(p_x^2 + p_y^2 + p_z^2\right) \tag{1.72}$$

Making the correspondence

$$p_x \to x \qquad p_y \to y \qquad p_z \to z \qquad E/c \to ct$$

we see that the structure of equation (1.72) is identical to that of the equation (1.44)

$$(ct)^2 - \left(x^2 + y^2 + z^2\right) = (ct')^2 + \left(x'^2 + y'^2 + z'^2\right)$$

which relates the respective space and time coordinates of the two observers, S and S′, and which was derived by simple algebra directly from the equations of the Lorentz transformation.

Hence, by analogy with equations (1.21), (1.23), (1.25) and (1.27) of the Lorentz transformation, the transformation equations of the components of the particle's linear momentum and total energy must be:

$$p'_{x'} = \gamma\left(p_x - uE/c^2\right) \qquad p'_{y'} = p_y \qquad p'_{z'} = p_z \qquad E' = \gamma\left(E - up_x\right) \tag{1.73}$$

Just as the four spacetime coordinates, x, *y*, z and t**,** can be depicted in a four dimensional representative space, so too can the four associated quantities p_x, p_y, p_z and E**.**

Sets of quantities which transform together in this way are called *four-vectors*.

Expressing physical laws in terms of four-vectors ensures their covariance under special relativity.

1.3.9 Magnetism - A Relativistic Effect

The application of special relativity to the field of electrodynamics results in an entirely new way of dealing with the magnetic phenomena associated with moving charges, one which is implicitly covariant and avoids the *ad hoc* postulates upon which Maxwell had based his theory.

To illustrate how this is done, we return to the case-study of Coulomb's Law and Magnetic Forces we first discussed earlier.

As we noted, whereas an observer at rest relative to a pair of identical positive charges finds the magnitude of the repulsive force acting between them to correspond to that given by Coulomb's Law, $F_E = kq^2/r^2$, a second inertial observer, relative to whom both the first observer and the positive charges are in uniform linear motion, finds that a weaker force, $F' < F_E$, acts between the charges.

Seemingly, Coulomb's law is not valid when the observer and the charges are in relative motion.

Thus, it should serve as a means for distinguishing between stationary and moving observers.

Classical electromagnetic resolves the discrepancy by the *ad hoc* postulate that each moving charges sets up a magnetic field, B, which, in turn, acts on the other such that an attractive force, $F_M = F_E - F' = Bqu$, acts between them.

However, as we noted at the time, this leaves the issue of the apparent nonequivalence of the inertial observers under Coulomb's law unresolved, for the magnetic postulate appears to be no more than an arbitrary contrivance designed to circumvent this.

Proper observer equivalence requires that the difference between the electric forces they record be accounted for in terms that have the same import for both, such as their relative motion.

Under the classical conception of motion, in which actual forces are invariants, this cannot be done.

We will now show how it is achieved under special relativity.

Consider a point charge, q, and a uniform line of charge, Q, of charge density λ (a long thin charged conductor) situated at a distance r from it.

In general, the coulombic force that acts between the charge and the line of charge is given by

$$F = \lambda q / 2\pi\varepsilon_0 r \tag{1.74}$$

and ε_0 is the vacuum permittivity.

The two inertial observers, S and S', measure the force that acts between the point charge and the uniform line of charge under three different circumstances:

(a) The point charge and the line of charge are at rest relative to the observer S.

(b) The point charge is at rest but the line of charge moves with a constant velocity u parallel to the X-axis relative to the observer S.

(c) Both the point charge and the line of charge move with a constant velocity u relative to the observer S.

(a) In this case (Fig 1.32a), the charge density S measures is the proper density, $\lambda_0 = Q/l_0$, and the force he observes acting parallel to the Y-axis is the proper force, $F_0 = \dfrac{\lambda_0 q}{2\pi\varepsilon_0 r}$. In special relativity, the corresponding force, F', measured by the moving inertial observer S' is obtained from the force transformation—equation (1.59): $F'_y = F_y/\gamma$. This gives $F' = F_0/\gamma$.

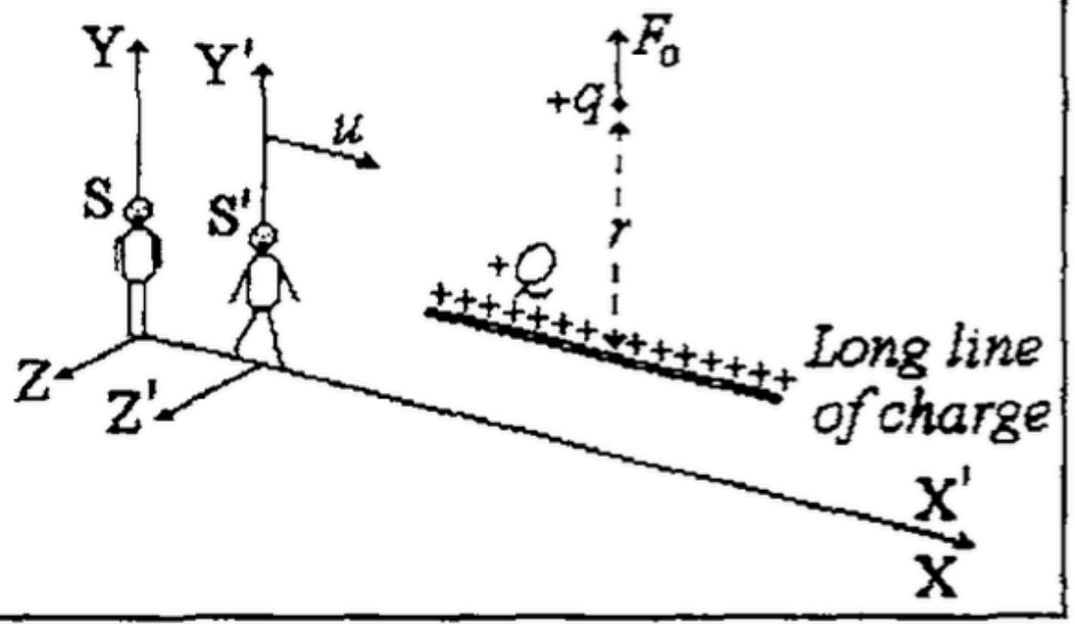


Fig 1.32a Relative to the observer S, the force acting on the point charge $+q$ is the proper force F_0. However, because of his motion relative to the point charge and line of charge, the observer S' measures a smaller force, $F' = F_0/\gamma$.

(b) In this instance, the observer S sees the line of charge moving parallel to the X-axis at a velocity u (Fig 1.32b). Consequently, relative to him, the line of charges contracts by a factor γ. This increase the charge density by the same factor making $\lambda = \gamma\lambda_0$. Hence the force he observes acting on the point charge is $F = \gamma F_0$.

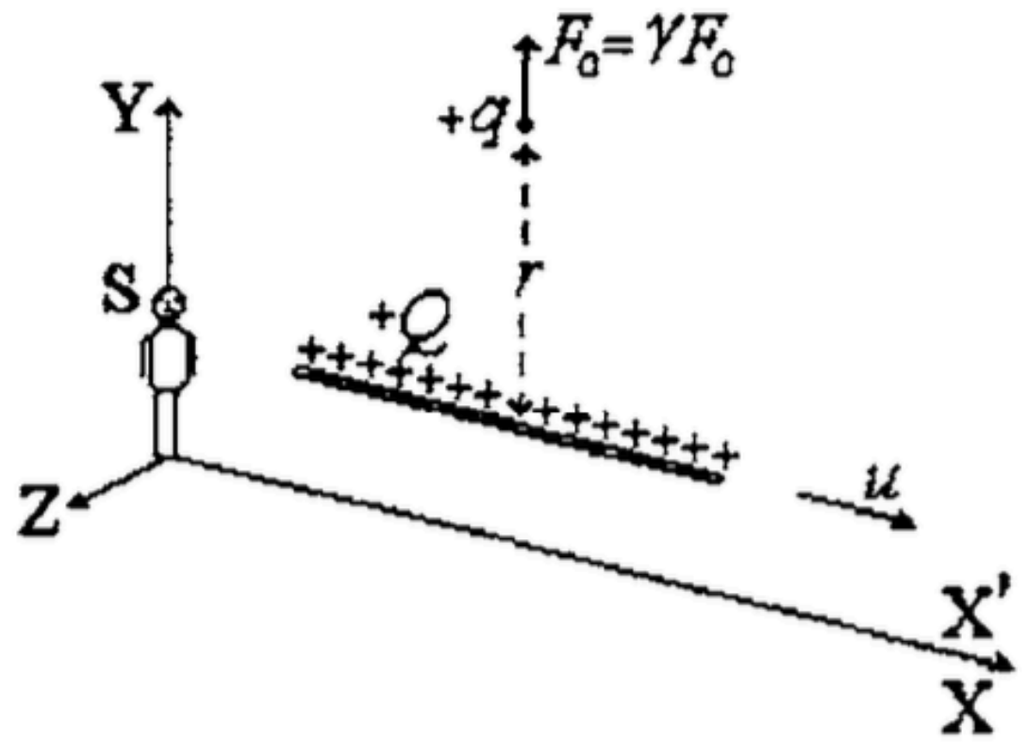


Fig 1.32b Relative to S, the length of the moving line of charge is foreshortened to $l = l_0/\gamma$.

The motion of the uniform line of charge is equivalent to a flow of electric charge—an electric current: $I = \lambda u$.

(c) As Fig 1.32c shows, except for the reversal in the direction of the relative motion, the stationary observer's view of the line of charge and of the point charge in this case is identical to that of the moving inertial observer, S', in case (a). Thus, the force measured by the stationary observer S in this case will be the same as that measured by S' in case (a), viz. F_0/γ.

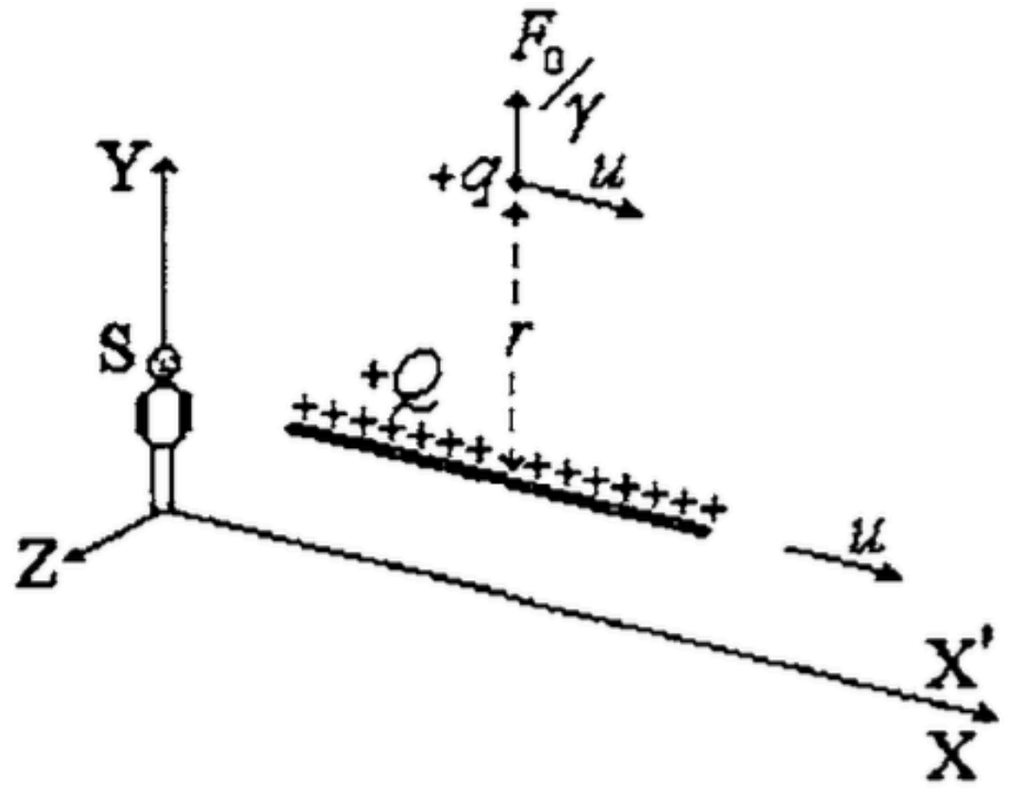


Fig 1.32c Both the point charge and the line of charge are in motion relative to S.

This case is equivalent to the familiar example of the force exerted by an electric current on a moving charge.

The difference, $(F_0/\gamma - \gamma F_0)$,between the forces acting on the point charge in cases (c) and (b) is equivalent to an attractive force acting towards and at right angles to the line of charge.

Under special relativity, this force, which is attributed to magnetic effects in classical theory, is simply a direct result of the motion of the point charge relative to S in case (c).

No further ad hoc postdates are needed in order to account for the difference between the forces: Coulomb's law and special relativity are all that is needed.

In this analysis, the reference frames of the two inertial observers, S and S′, are symmetrical and interchangeable.

Thus, if Coulomb's law and special relativity are all that is required to account for the forces observed by S, the same must be true for S′.

Hence, S and S′ are equivalent and Coulomb's law is covariant.

In general, any difference in the electrical forces inertial observers measure results only from their relative motion; there is no need for the magnetic force postulate.

Under special relativity, what are usually called magnetic forces arise naturally from the motion of electrical charges; they are not a distinct or separate phenomena but simply the velocity dependent component of the electrical interaction between the charges.

Why then does electromagnetic theory work so well? Why can classical electromagnetism with its *ad hoc* postulates still be used?

The answer is that making the Maxwellian magnetic postulate is in fact tantamount to adopting the Einsteinian conception of motion when employing Coulomb's Law.

To put this another way: the relativistic force of attraction, $(F_0/\gamma - \gamma F_0)$ that acts on the point charge is actually identical to the magnetic force, $F_M = Bqu$, postulated by classical electromagnetic theory.

That this is so, is readily demonstrated by the following series of substitutions

Putting $F_M =$ $(F_0/\gamma - \gamma F_0)$ and rearranging gives

$$F_M = \gamma F_0 \left(\frac{1}{\gamma^2} - 1 \right) \tag{1.75}$$

Substituting $F_0 = \dfrac{\lambda_0 q}{2\pi\varepsilon_0 r}$, followed by $\lambda = \gamma\lambda_0$ and $\gamma = \sqrt{\frac{1}{1-u^2/c^2}}$ yields:

$$\begin{aligned} F_M &= \frac{\gamma\lambda_0 q}{2\pi\varepsilon_0 r}\left(\frac{1}{\gamma^2} - 1 \right) \\ &= \frac{-\lambda u^2 q}{2\pi\varepsilon_0 r c^2} \end{aligned} \tag{1.76}$$

The negative sign indicates that the force is one of attraction.

Finally, incorporating the electromagnetic definition of the speed of light, $c = \sqrt{\frac{1}{\varepsilon_0\mu_0}}$ in equation (1.76) produces:

$$|F_M| = \frac{\mu_0 \lambda u}{2\pi r} \cdot qu \tag{1.77}$$

Relative to the observer S, the motion of the line of charge in both (b) and (c) is equivalent to a current, $I = \lambda u$.

Making this further substitution in equation (1.77), we obtain the familiar electromagnetic expression

$$|F_M| = \frac{\mu_0 I}{2\pi r} \cdot qu = Bqu \tag{1.78}$$

for the force acting on a point charge, q, moving with velocity u parallel to a straight conductor carrying a current I, at a distance r from it.

To summarize.

Whether we introduce the magnetic field by making the magnetic postulate, i.e., that a moving or changing electric field sets up a magnetic field, or by combining Coulomb's law and the Einsteinian conception of motion, is really just a matter of convenience.

One approach is no more fundamental than the other.

The same reality underlies Faraday's fields, Maxwell's equations and special relativity.

Which way we choose to approach this reality is a matter of individual choice.

Part 1.4 The General Theory of Relativity

The jewel in the crown of classical mechanics was *Newton's Theory of Gravitation*.

Taken together with his laws of motion, it afforded explanations and predictions of unprecedented accuracy, that were repeatedly confirmed by experiment and observation.

The classic example was Newton's own derivation and theoretical elucidation of Kepler's three empirical laws.

However, for all its success, classical gravitational theory gave no clear explanation of the source of the force that attracted one body to another.

Nor did it explain how this force acted across space: instantaneously or through some medium.

Newton himself was circumspect on all these points.

Though he recognized that a body's weight is directly proportional to its inertia, its immutable *vis inertia*, he rejected the notion that "gravity is essential to bodies".

As he wrote:

> *It is inconceivable, that inanimate brute matter: should, without the mediation of something else, which is not material, operate upon and affect other matter without mutual contact, as it must be, if gravitation be essential and inherent in it…That gravity should be innate, inherent and essential in matter: so that one body may act on another at a 'distance through a vacuum, without the mediation of anything else, is to me so great an absurdity. that I believe no man, who has a competent faculty of thinking, can ever fall into it.*

But if gravity is not an inherent property of matter, then of what is it a property?

Newton refused to offer any answer to this question.

> *Hitherto we have explained the phenomena of the heavens and of our sea by the power of gravity, but have not yet assigned the cause of this power…and I frame no hypotheses.*

Others were less reticent.

Notwithstanding Newton's clear rejection of the notion, even during his lifetime, his theories were interpreted as implying that gravity is indeed an innate inherent property of matter.

In the Preface he was asked to write in 1713 to the second edition of *Principia,* the astronomer Cotes asserted that "the attribute of gravity is found in all bodies".

A corollary of this interpretation was that *actions at a distance,* such as gravitational interactions, propagate instantaneously across space.

However, with the establishment of the theory of special relativity, which prohibits instantaneous interactions of this type, this became untenable.

Ironically, it was the very accuracy of its predictions and the improvements in instrumentation that they motivated, that gave rise to the first suspicions that something was not quite right in Newton's theory of gravitation.

Towards the end of the 19th century, it was found that even after allowing for all the perturbations in their orbits that result from their mutual attractions and from those of their moons, the orbits of the planets deviate very slightly from those predicted by Newton's law of gravity.

The effect is most pronounced in the case of Mercury, the planet closest to the sun.

According to Kepler's Laws, at the end of each planetary year, a planet should return to the same point in space it occupied at the beginning of the year.

However, the perihelion of the planet Mercury-the point closest to the sun in the planet's orbit-is not a fixed point in space; the perihelion precesses.

In other words, at the end of each planetary year, Mercury overshoots the point in space it occupied at the beginning of the year.

The full extent of this precessional motion is inexplicable in Newton's theory of gravitation (Fig 1.33).

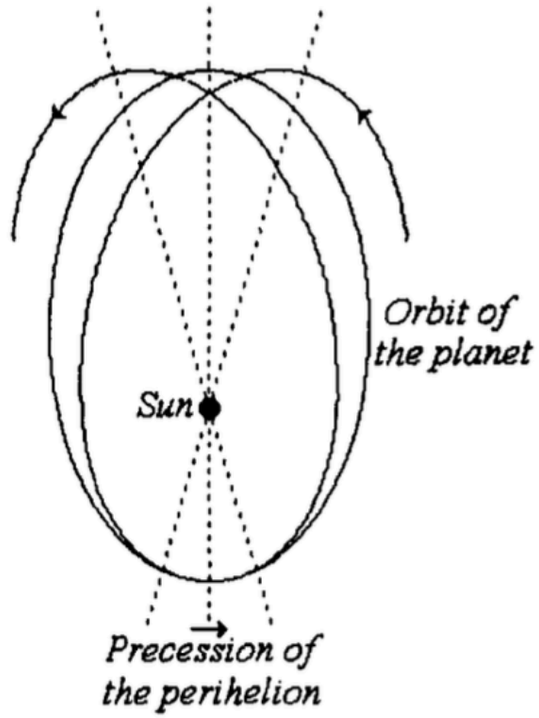


Fig 1.33 The precession of the perihelion of a planet. After each orbit the perihelion advances in the direction of the planet's motion to a new point in space. The direction of the major axis of Mercury's orbit changes by about 43 seconds of arc per century.

1.4.1 The Postulates of the General Theory of Relativity

The precession of the perihelion of Mercury's orbit was first explained by Einstein when he broadened his special theory of relativity, which had dealt only with inertial observers, into a general theory that pertained to all observers, irrespective of their coordinate system.

Accordingly, the new *Principle of General Relativity* that he formulated was:

> **The laws of physics must be expressed in such a way that they are form invariant in all reference frames, whatever their motion.**

The exercise of this principle requires that there be a transformation under which all the laws of physics are covariant, whether or not the reference frames from which they are formulated are inertial.

This transformation must do for all observers, irrespective of their motion - accelerated or otherwise - what the Lorentz transformation does for inertial observers.

In particular, it must reconcile any differences between the observations made in the various reference frames, such that no observer should be able to determine, by means of any experiment, whether he is at rest; in uniform motion in a straight line or in an accelerated motion.

At first sight, it would appear that our daily experiences belie this corollary of the principle of general relativity.

Based on the inertial forces we feel at the time, we believe we can always tell whether or not we are accelerating.

The palpable forces we experience when traveling in a car or a bus at every turn, at every bump and at every time it stops or starts, tell us that we are in an accelerating non-inertial reference frame.

Such forces are not felt by passengers in a stationary vehicle nor by those in a vehicle that is moving along a perfectly straight and smooth horizontal road at uniform velocity.

In classical mechanics these inertial or fictitious forces serve to distinguish, absolutely, between inertial and non-inertial frames of reference.

Newton regarded inertial forces, which do not depend on interactions between bodies and which appear to be a universal phenomenon, as providing the evidence for the existence of *Absolute Space*. He wrote:

> *For instance, if two globes kept at a given distance one from the other, by means of a cord that connects them, were revolved about their common centre of gravity; we might, from the tension of the cord, discover the endeavor [tendency] of the globes to recede from the axis of their motion and from this we might compute the quantity of their circular motion even in an immense vacuum, where there was nothing external or sensible with which the globes could be compared.*

Newton argued that the tension in the cord linking the two rotating globes was an intrinsic attribute of the system of rotating spheres.

Its origin was in their rotation and in this alone.

Thus it constituted absolute proof of their rotation without the need for any external point of reference.

Space itself was the frame for their rotation.

It is only on this basis that the Copernican heliocentric system can be accorded preference over the Ptolemaic geocentric system.

The description of planetary motions given by the Copernican system is not inherently superior to that given by the Ptolemaic.

For it makes no difference to our view of the heavens whether the earth spins on its axis once every 24 hours as is postulated by the Copernican system or the heavenly vault orbits a stationary earth once every 24 hours as the Ptolemaic system posits.

Under the Newtonian conception, the real difference between the two models is that if the earth was at rest, free bodies would move across its surface in straight lines.

But if it is spinning, their trajectories should be curved - the *Coriolis eflect.*

Satellite photographs such as those seen every evening in television weather forecasts show masses of atmospheric air moving across the earth's surface in curved paths.

Under the Newtonian conception, this proves that the earth is not at rest; that inertial forces arise in the atmosphere and determine the movement of the cyclones and anti-cyclones associated with terrestrial weather patterns.

These forces would not arise if the earth was at rest.

Classically, it is the weather that proves the earth is spinning.

However, we may ask whether this is the only explanation that can be given for these terrestrial effects?

Are changes in motion-accelerations-the only possible way of accounting for the phenomena?

Does the presence or absence of inertial frames unequivocally distinguish **between** non-inertial and inertial reference frames?

Suppose, that our vehicle runs along a perfectly smooth horizontal track which holds it upright at all times and that its windows are all covered so that we cannot look out.

Initially, the vehicle cruises at a uniform speed (there is a speedometer inside) along what we feel to be a straight stretch of track.

Suddenly, we and all the other passengers and objects in the vehicle are flung sideways, to our left (Fig 1.34).

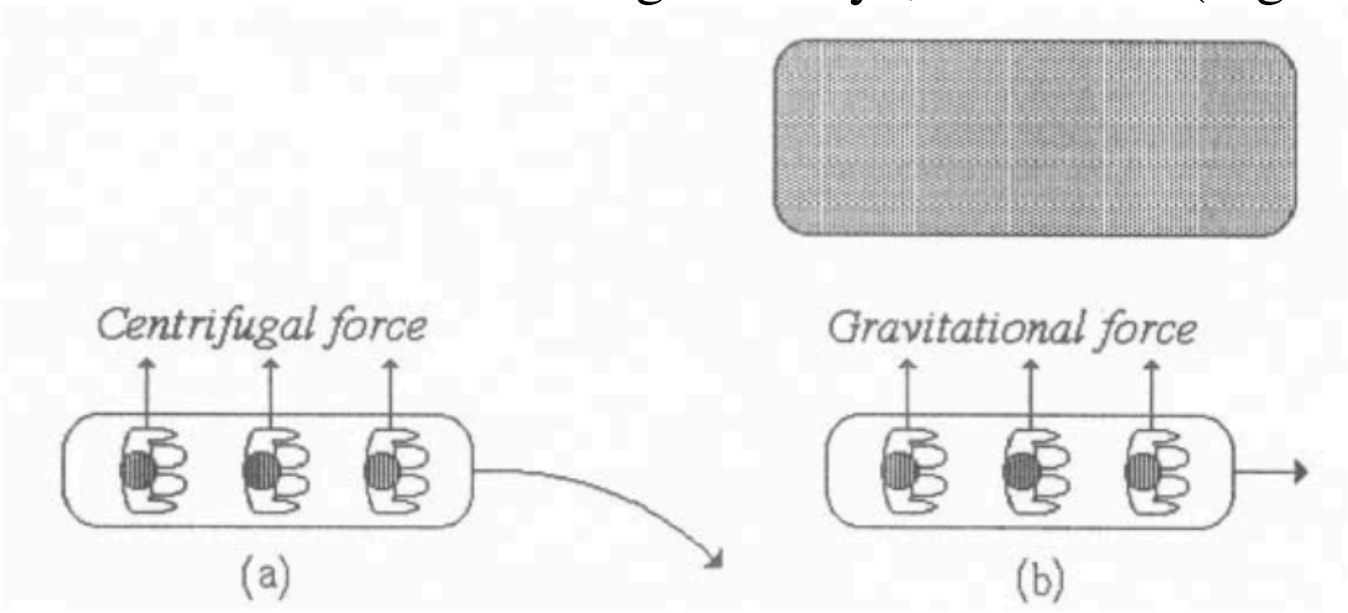


Fig 1.34 The passengers in the vehicle cannot distinguish between the two possibilities: (a) The vehicle made a sharp right turn and as a result a centrifugal (inertial) force acted on them throwing them to their left; (b) A massive body, whose gravitational pull drew the passengers towards it, suddenly appeared to the left of the vehicle.

According to the reasoning above, we conclude that the vehicle had come to a sharp bend in the track, where, without slowing down, it made a right turn.

As a result of this change in direction, an inertial force-a *centrifugal* force-had acted upon us, throwing us to the left.

However, there is an alternative explanation to the sensations we experience, one that does not involve any change in the our motion.

It is that a massive body, which exerts a substantial gravitational pull, suddenly appeared for a few moments outside and to the left of the vehicle, and that we had all ‘fallen’ sideways towards it because of our own gravitational mass.

The vehicle had in fact continued to move in a straight line; there was no bend in the track.

Though we may consider this second explanation to be somewhat far-fetched, it cannot be dismissed out of hand.

Einstein argued, that the two explanations are equally valid and, moreover, that they are indistinguishable.

Consequently, there is no way we can know, absolutely, why we were flung sideways.

It might have been because of the inertial forces generated by a change in the direction of our motion; equally, it might have been the result of a gravitational force as we continued to travel in a straight line.

In the former case, there would have been an acceleration; in the second, there would not.

Thus, contrary to our intuition, we cannot always tell whether or not we are accelerating; it may be gravity.

Einstein summarized this notion in a postulate known as the *Principle **of** Equivalence*:

> **An observer cannot determine, in any way whatsoever, whether the laboratory he occupies is in a uniform gravitational field or is in a reference frame that is accelerating relative to an inertial frame.**

According to this principle, all the phenomena experienced by a body in a gravitational field can be duplicated by means of an appropriate acceleration.

And, *vice versa*, all those experienced by an accelerated body can be duplicated in an appropriate gravitational field (Fig 1.35).

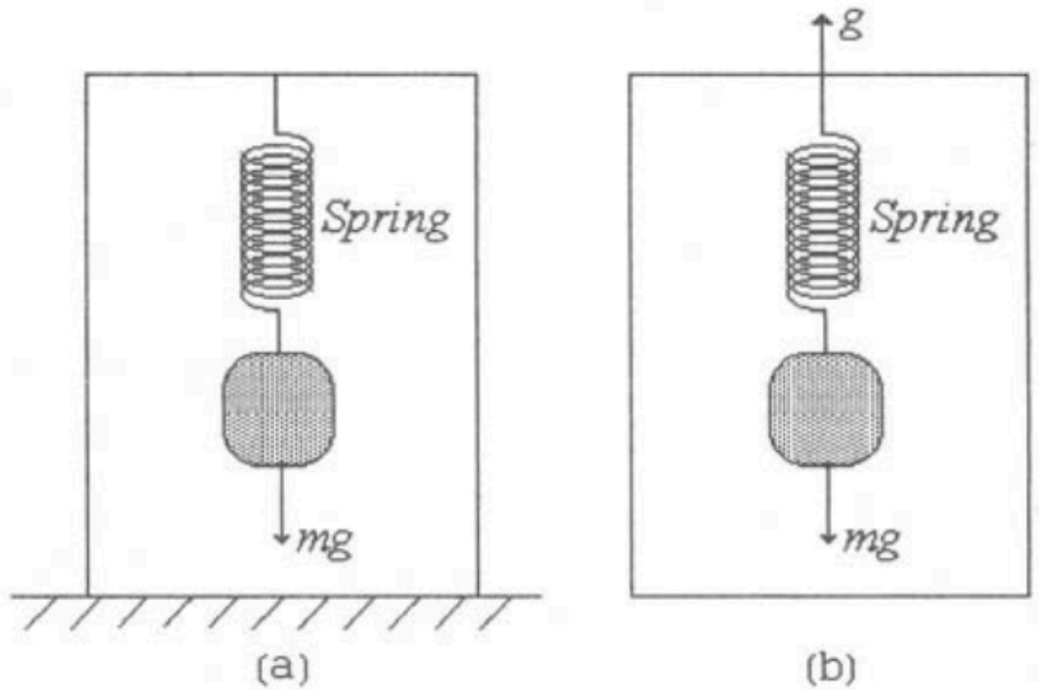


Fig 1.35 Two equivalent elevators: (a) an elevator at rest on the earth's surface; (b) an elevator in deep space rising with an acceleration g. Identical objects hang from identical springs, each of which is stretched to the same degree. Each duplicates the physical properties of the other. Observers inside the elevators cannot distinguish between the two possibilities.

Consequently, by choosing the appropriate acceleration for our reference system, we can duplicate the effects of a constant gravitational field, reduce or increase the strength of a given field, or even neutralize it completely.

The notion underlying the principle of equivalence is the sameness of gravitational mass and inertial mass.

According to the general theory of relativity, gravitational and inertial mass are not different properties of matter but two aspects of a fundamental and single property of matter.

The equality of the magnitude of an object's gravitational mass with that of its inertial mass, which, in classical physics, explains why all bodies fall to the earth's surface with the same acceleration, g, whatever their size or composition, is not fortuitous; it reflects an underlying fundamental unity.

It must be emphasized, that the principle of equivalence is valid only for uniform gravitational fields or for infinitesimal volume elements of non-uniform fields.

The reason for this is that an observer in a non-uniform gravitational field could detect its presence by releasing two objects.

Due to the non-uniformity of the field the objects would either move towards each other or away from one another.

We say that tidal forces act in a non-uniform gravitational field.

There are no tidal forces in a uniform gravitational field nor can they be detected in an infinitesimal volume element of a non-uniform field.

Thus, the field in an infinitesimal volume element of any gravitational field can be considered, at least for short periods of time, to be uniform.

Consequently, we can always posit an accelerated reference frame, relative to which, in a small region of spacetime, the gravitational field is cancelled out.

1.4.2 Gravitation and the General Theory of Relativity

Between the ***years*** 1907 and 1916**,** Einstein succeeded in developing a theory of gravitation that is valid for all observers whatever their motion; a truly covariant theory.

To a first approximation,Newton's law of gravity ensues from it as the case of the gravitational field around a point mass.

In the case of weak gravitational fields, Einstein's theory and that of Newton reach the same conclusions.

A full elucidation of Einstein's theory of gravity requires the solution of the field equations he developed, a task that demands complex mathematical processes and is far beyond the intent of this discussion**.**

However, some of the physical consequences of the theory, such as the effect of gravitational fields on the measurement of physical quantities, can be deduced from a combination of the principle of equivalence and the special theory of relativity.

This is the approach Einstein himself took when he first predicted some of the more surprising consequences of his theory.

To ensure the validity of the use of special relativity, we must restrict ourselves to the short distances and short periods of time in which the motions can be considered to be uniform.

As an example of how gravity can be neutralized, consider the following.

Suppose, that the cable holding a windowless elevator at rest close to the earth's surface snaps.

As a result, the elevator and all its contents fall freely in the earth's gravitational field.

An observer inside the lift would experience the sensation of weightlessness; he would find no evidence of the gravitational field.

The acceleration of its free-fall neutralizes the effect of the gravitational field inside the elevator.

Unable to look out, the observer might well believe that he had been somehow transported to a *gravityless* inertial reference system, deep in space far from any gravitational field.

We may summarize this as follows:

Experimentally, a small free-falling laboratory is equivalent to an inertial reference system in which there is no gravitational field.

Hence, the effect of a gravitational field on the measurement of a physical quantity can be investigated by comparing its measurement by an observer in a small free-falling laboratory, with that made by a stationary observer at the same point in the gravitational field.

Suppose, that a stationary observer, situated at a point above the earth's surface and at a distance r from its the center, measures (i) the time period between events occurring in his neighborhood and (ii) the lengths, parallel to the direction of the gravitational field, of objects at rest around him.

A second observer, who has fallen freely in a small laboratory from rest at infinity, is momentarily at the same distance from the earth's center, close to the stationary observer.

The free-falling observer measures the time period between the same events as the stationary observer and the lengths, in the direction of his motion, of the same objects (Fig 1.36).

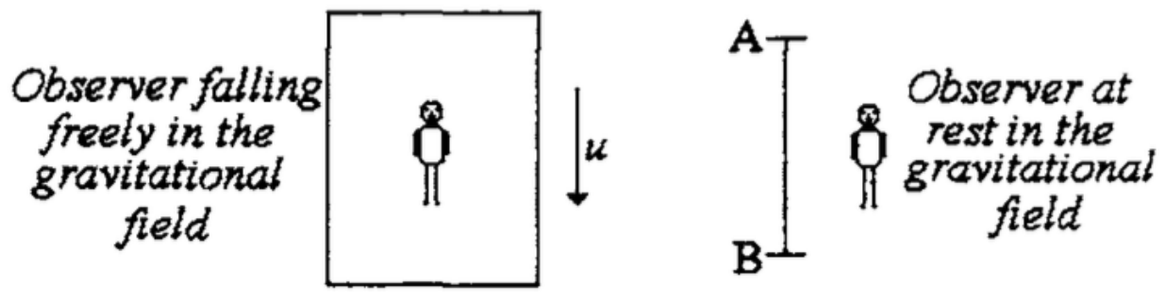


Fig 1.36 Two observers in a gravitational field, one stationary and the other free-falling. Whereas the stationary observer feels the effects of the field, the one in the free-falling laboratory experiences weightlessness, his situation is equivalent to that of an observer in a region where there is no gravitational field. Both observers measure the length AB of an object at rest relative to the stationary observer.

In terms of the special theory of relativity, the time intervals and lengths measured by the stationary observer, Δt_0 and L_0, will be proper time intervals and proper lengths respectively, for he is at rest relative to the measurements.

Thus, the corresponding time internals, Δt, measured by the free-falling observer will be longer than those measured by the stationary observer and the corresponding lengths, L, that he measures will be shorter than those measured by the stationary observer.

The proper quantities measured by the stationary observer are related to the corresponding quantities measured by the free-falling observer by the equations $L_0 = \gamma L$ and $\Delta t = \gamma \Delta t_0$, respectively.

The factor $\gamma = \sqrt{\frac{1}{1-u^2/c^2}}$ where u is the instantaneous relative velocity of the two observers

The instantaneous relative velocity of the two observers, u, is the speed of descent of the free-falling laboratory.

At infinity, the total potential and kinetic energy of the small laboratory was zero; at the point above the earth's surface where it passes by the stationary observer, its total potential and kinetic energy is

$$-GMm/r + mu^2/2$$

Conservation of the energy gives $-GMm/r + mu^2/2 = 0$ from which

$$u = \sqrt{2GM/r} \tag{1.79}$$

where m is the mass of the laboratory, M is the mass of the earth and G is Newton's universal gravitational constant.

Substituting $u = \sqrt{2GM/r}$ in $\gamma = \sqrt{\frac{1}{1 - u^2/c^2}}$ gives:

$$\gamma = \sqrt{\frac{1}{1 - 2GM/rc^2}} \tag{1.80}$$

Hence, the closer the observers are to the earth's center, the larger the factor γ and the greater will be the differences between the measurements they make.

In general, the stronger the gravitational field or, equivalently, the greater the acceleration of the moving observer, the greater the difference between the measurements made by the stationary and free-falling observers.

The minimum value that r can take, $r_S = 2GM/c^2$, namely, that for which y = ∞ , is called the *Schwarzschild radius*.

It is the critical radius into which a sphere of mass M must be compressed to form a non-rotating black hole.

It specifies the boundary of the black hole-the event horizon-inside which all matter and radiation are drawn into the singularity at its center where the known laws of physics break down.

A corollary of the effect of gravity on time measurements is that an observer in a weak gravitational field should measure a lower vibrational frequency than that measured by an observer in a strong gravitational field.

For instance, because of the non-uniformity of the earth's gravitational field, spectral lines should be red-shifted (displaced towards the lower frequencies) in a laboratory on the top floor of a building as compared to one on the ground floor.

The difference is minute but it was detected in 1960 with the help of the Mossbauer effect.

Apart from its effect on measurements of space and time, the gravitational field also affects mass; an object's mass increases, relative to an observer in a gravityless laboratory, the stronger the gravitational field in which it is situated.

Consequently, mass cannot be defined without reference to gravity.

The fact that measurements of spatial displacements and of time intervals **are** affected by the gravitational field in which they are made, provides the first hint of the most radical change brought about by Einstein's theory of general relativity: the notion that gravity is something fundamentally geometrical.

1.4.3 Gravity and Geometry

In its elucidation of the motion of bodies, classical mechanics is totally reliant on the concept of force: whether by virtue of their absence as in inertial reference frames; their presence as when an actual physical interaction takes place; or their mystical summoning-up from absolute space, as though by fiat, *ex nihilo*, in non-inertial **systems.**

Whatever the object's state of motion-at rest, in uniform motion in a straight line or an accelerated motion - forces are invoked in accounting for it. Consider the following three instances:

a) **A weightless observer throws an object across his gravityless laboratory, from one side to the other. The object moves in a straight trajectory, no external forces being exerted on it during its flight. The laboratory is an inertial system in which Newton's laws of motion are obeyed (Fig 1.37a).**

b) **An observer in an accelerated laboratory, deep in space far from any gravitational field, also throws an object across his laboratory, launching it at right-angles to the direction of his acceleration. Despite the fact that here too, no actual external force was impressed upon it, the object follows a parabolic path. In classical mechanics, the phenomenon is explained by the contrivance of an inertial force—a fictitious force—that is somehow exerted on the object by virtue of it being situated in an accelerated laboratory; it's all a bit metaphysical (Fig 1.37b).**

c) **A third observer, at rest in a uniform gravitational field, also throws an object across his laboratory, launching it at right-angles to the direction of the field. This object also moves in a parabolic trajectory. To explain this within the framework of his laws of motion, Newton invented a new interaction: gravity (Fig 1.37c). However, according to the principle of equivalence, gravitational effects are indistinguishable from those of an acceleration. Are we therefore to conclude, by analogy, that the force of gravity is a metaphysical contrivance too?**

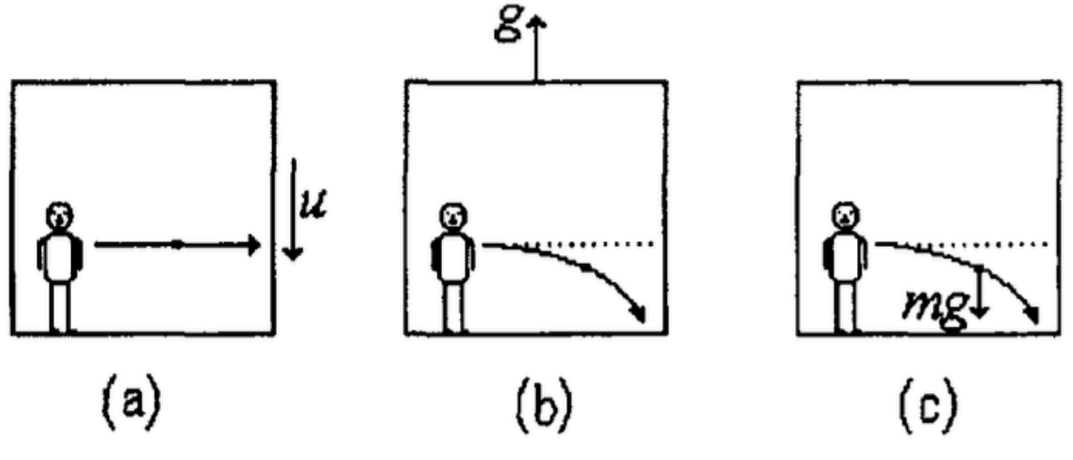


Fig 1.37 Three observers performing the same experiment. (a) A weightless observer throws an object across his free-falling laboratory; its instantaneous velocity is u. (b) An observer throws an object across his laboratory as it is rises with acceleration g. (c) An observer throws an object across his stationary laboratory in a uniform gravitational field of strength g.

General relativity requires that all observers, whether they be in inertial or accelerated reference frames or in gravitational fields, be indistinguishable.

However, in the instances described above, it would appear that the first observer can be distinguished from the other two by virtue of the shape of the object's trajectory.

Whereas he finds the trajectory to be a straight line, the others find it to be a parabola.

In order for the three observers to be indistinguishable, the straight trajectory observed in the inertial frame must be the same for that observer as is the parabolic trajectory observed by the observers in the non-inertial frames.

Einstein realized that, as long as the laws of motion were expressed in terms of forces, this could never be so; that for these laws to be truly invariant, they had to involve no more than just the relative positions and motions of the bodies.

Accordingly, he restated Newton's First Law of Motion as follows:

Every object, left alone, moves in a straight line.

At first sight, this seems to solve nothing, for the trajectories observed by the non-inertial observers still appear to us to be parabolas; they do not look like straight lines.

Einstein countered this by noting that we had all been taught to believe, from a young age, that the flat geometry associated with the name of Euclid and employed by Newton in his *System of the World*, correctly describes physical reality.

However, during the 19th century, mathematicians had questioned some of the axioms underlying Euclidean geometry and by amending one or more of them had developed exotic new curved geometries.

Moreover, it had been shown that the geometry of a surface could be expressed in terms of a mathematical tool called a tensor which, like a vector, is inherently invariant; vectors are in fact simple tensors.

Einstein suggested that if the tensor forms of these new non-Euclidean geometries were used to depict physical reality, it would be possible to transform the parabolas of the non-inertial observers into the *straight line* of the inertial observer as required by the principle of general relativity and so to describe motion without resort to forces, actual, inertial or even gravitational.

A simple way of distinguishing between the various geometries is in terms of the value they assign to the ratio between the circumference of a circle and its radius (Fig1.38).

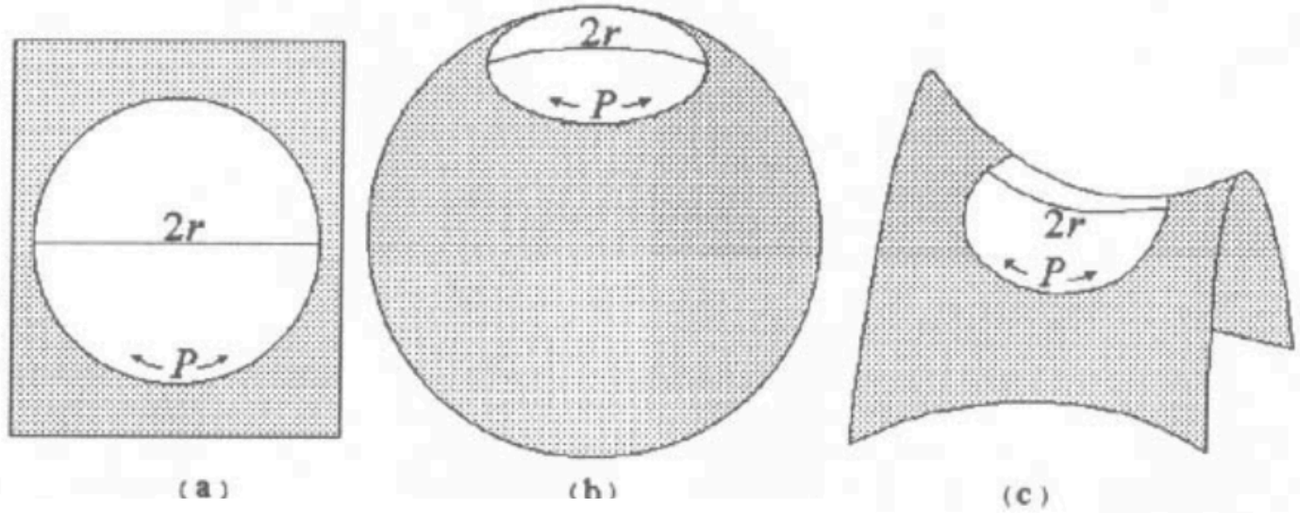


Fig 1.38 Three different geometries. (a) Euclidean: the ratio, P/r , between the circumference, P, and the radius, r, of a circle drawn on a flat surface is 2π. (b) Elliptical: for a circle drawn on a sphere $P/r < 2\pi$. (c) Hyperbolic:for a circle drawn on a saddle surface $P/r > 2\pi$.

* **For circles drawn on the conventional flat surfaces of Euclidean geometry, this ratio equals 2π.**
* **For circles drawn on the curved spherical surfaces of the so-called *elliptical geometry*, the ratio is less than 2π.**
* **For circles drawn on saddle shaped surfaces of *hyperbolic geometry*, the ratio is greater than 2π.**

In all geometries, the *straight line* is the line of shortest length that connects two points; mathematicians call these lines *geodesics* or *geodetic lines*.

Their shape depends on the particular geometry.

For example, the geodesics of Euclidean geometry are straight lines, whereas those of the elliptical geometry are the great circles on a spherical surface.

Geodesics are described by invariant formulae from which all the other invariants of the surface can be derived.

Consequently, they require no transformation when going from one coordinate system to another.

Though its shape may change, the line that is a geodesic in one reference frame will also be one in another.

Consequently, by describing motion exclusively in terms of geodesics, Einstein could formulate a new covariant mechanics and theory of gravity.

According to this new mechanics, the trajectories of all bodies through spacetime are geodesics.

This axiom is consistent with the formulation of Newton's first law of motion given above and enables us to express the laws of mechanics in terms of a real physical quantity-the shortest distance between two points-rather than the spurious concepts of inertial and gravitational forces.

Similarly, gravity is associated with the geometry of spacetime.

A body's presence affects the shape of the spacetime in its neighborhood and that of its geodesics.

This, in turn, affects the trajectories of other bodies moving through the spacetime.

The classical theory of gravitation is applicable in small spaces and in weak fields, where to a close approximation, Euclidean geometry is valid.

Suppose, that a circle is drawn around the center of the earth and two observers set out to measure its circumference and radius.

One observer is in a free-falling laboratory - a gravityless laboratory, and the other in a stationary laboratory in the earth's gravitational field (Fig 1.39).

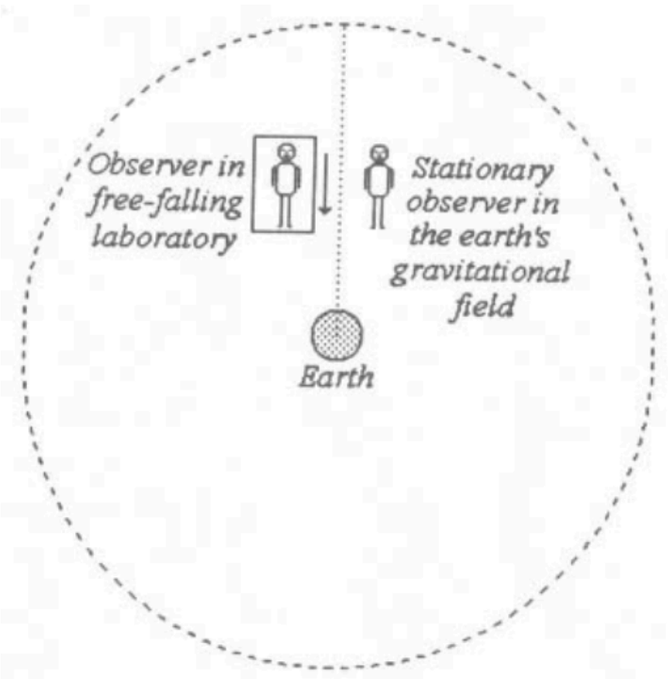


Fig 1.39 A circle drawn round the center of the earth. Relative to the falling observer, the circle's radius, r, is shorter than the proper radius, r_0, measured by the stationary observer. The two observers agree about the circle's circumference.

Since the gravitational field is perpendicular to the circle at all points, the two observers agree as to its circumference, P.

However, the proper radius, r_0, measured by the stationary observer is greater than the radius, r, measured by the free-falling observer.

The free-falling laboratory is equivalent to an inertial reference system in which Newton's laws and Euclidean geometry are valid.

Hence, the ratio, P/r , between the circumference of the circle and the radius measured by the free-falling observer is equal to 2π.

On the other hand, since $r_o > r$, the ratio, P/r_0, between the circumference of the circle and the radius measured by the stationary observer is less than 2π.

Thus, the geometry appropriate to the stationary observer in the earth's gravitational field is the elliptic geometry.

The gravitational field experienced by the stationary observer is a manifestation of the difference between his geometry and that of the free-falling observer.

It can be shown that an observer rotating on a carousel finds that the ratio between the circumference of the circle and its radius is greater than 2π; the geometry applicable to an observer in a rotating laboratory is therefore the hyperbolic geometry.

All the phenomena classically attributed to centrifugal forces and the Coriolis effect can be elucidated by describing the motions observed from the carousel in terms of geodesics in this non-Euclidean geometry.

We can now return to the precession of the perihelion of the planet Mercury.

Kepler's laws, like Newton's theory of gravitation, presume that Euclidean geometry is right for describing a planet's motion.

However, the planet Mercury is the closest of all to the sun and so it moves in an appreciable gravitational field.

The geometry that should be used to describe its motion is the elliptic geometry in which the circumference of a circle (or ellipse) is less than that of a circle (or ellipse) with the same radius in Euclidean space.

The periodic time (sidereal period), T, calculated from Kepler's laws is the time it takes the planet to return to a given point in its orbit relative to the sun.

In Euclidean geometry this is also the time it takes for the planet to return to a given point in space.

But this is not so in non-Euclidean geometry.

Because of the elliptic geometry of the space in which it moves, the planet Mercury has to travel a shorter distance than the length of its Keplerian orbit in order to return to a given point in space.

Consequently, when, after a time T, the planet returns to its perihelion - the point in its orbit closest to the sun - it has overshot the point in space that was the previous location of its perihelion.

As a result of this, the perihelion *precesses* through space in the direction of the motion of the planet.

Newtonian gravity obeys the principle of superposition, i.e., the field, g, that it predicts for any point is the sum of the fields produced there by all the masses in space: $g = g_1+g_2\ldots+g_n$.

But in the Einsteinian theory, the shape of spacetime which determines the gravitational field at a point, is not that predicted by simply adding together the effects of the different masses in space on the geometry at the point.

In general relativity, not only do bodies affect the shape of spacetime, but their separate effects also act on one another and contribute to the final field.

This interaction affects the energy of the gravitational field and is the source of the physical difference between the classical and the relativistic fields.

In the Einsteinian universe, changes in the shape of spacetime propagate with the speed of light; waving our hands does not instantaneously move the galaxies.

The changes in the shape of spacetime are accompanied by energy changes which, it has been suggested, should be observable as *gravity waves*.

However, gravitational changes take place over aeons and **so** such waves will be of very low intensity, which makes their detection very difficult.

The world-lines of light rays are geodesics too, whose shape depends on the gravitational field through which the light propagates.

Hence, in the narrow Euclidean sense, light does not propagate in a straight line in a gravitational field; light rays cannot be used to draw orthogonal coordinate systems in which Euclidean geometry is valid.

Time intervals and lengths too depend on the gravitational fieids in which they are measured.

Consequently, it will come as no surprise to learn that according to general relativity, the velocity of light in *vacuo* depends on the strength of the gravitational field through which it propagates.

The speed of light in *vacuo* is not the supreme law of Nature that special relativity supposes it to be.

As Einstein wrote:

> *The principle of the constancy of the velocity of light in vacuo, which is one of the two postulates of the special theory of relativity, does not have an unlimited validity. Light rays can only bend if the velocity of light changes from place to place ... The results of special relativity are valid only so long as the effects of gravitational fields on the phenomena can be ignored.*

Now we finish Part 1 with some detailed derivations of special relativity properties.

1.5.1 Velocity Addition in Special Relativity

Consider an object moving relative to the two observers, S and S′. The components of its velocity relative to the observer S are

$$v_x = \frac{dx}{dt} \qquad v_y = \frac{dy}{dt} \qquad v_z = \frac{dz}{dt} \tag{1.81}$$

and relative to the observer S′

$$v'_{x'} = \frac{dx'}{dt'} \qquad v'_{y'} = \frac{dy'}{dt'} \qquad v'_{z'} = \frac{dz'}{dt'} \tag{1.82}$$

In general, the velocity components measured by S can be related to the corresponding components measured by S′, by making appropriate substitutions in the above definitions. For instance, to express the S components in terms of the S′ components, expressions for the infinitesimals dx, dy, dz and dt in terms of dx', dy', dz' and dt' are substituted in their place in the S definitions. The relevant expressions are obtained from the derivatives of the equations of the Lorentz transformation.

$$dx' = \gamma(dx - udt) \qquad dx = \gamma(dx' + udt') \tag{1.83}$$

$$dy' = dy \qquad dy = dy'$$

$$dz' = dz \qquad dz = dz'$$

$$dt' = \gamma\left(dt - {udx}/{c^2}\right) \qquad dt = \gamma\left(dt' + {udx'}/{c^2}\right)$$

For example, to obtain an expression for the component $v'_{x'}$ in terms of v_x:

$$v'_{x'} = \frac{dx'}{dt'} = \frac{\gamma(dx - udt)}{\gamma\left(dt - {udx}/{c^2}\right)} \tag{1.84}$$

and substituting $dx = v_x dt$ from equation (1.81) gives

$$v'_{x'} = \frac{\gamma\left(v_x dt - u dt\right)}{\gamma\left(dt - {uv_x dt}/{c^2}\right)} \tag{1.85}$$

$$= \frac{v_x - u}{1 - {uv_x}/{c^2}}$$

Similarly for the other two velocity components:

$$v'_{y'} = \frac{dy'}{dt'} \tag{1.86}$$

$$= \frac{dy}{\gamma\left(dt - {uv_x}/{c^2}\right)}$$

$$= \frac{v_y}{\gamma\left(1 - {uv_x}/{c^2}\right)}$$

and

$$v'_{z'} = \frac{v_z}{\gamma\left(1 - {uv_x}/{c^2}\right)} \tag{1.87}$$

The corresponding equations which give the velocity components relative to the observer S in terms of those of S′ are:

$$v_x = \frac{v'_{x'} + u}{1 + {uv'_{x'}}/{c^2}} \tag{1.88}$$

$$v_y = \frac{v'_{y'}}{\gamma\left(1 + {uv'_{x'}}/{c^2}\right)} \tag{1.89}$$

$$v_z = \frac{v'_{z'}}{\gamma\left(1 + {uv'_{x'}}/{c^2}\right)} \tag{1.90}$$

Equations (1.85) to (1.90) are the velocity addition rules of the special theory of relativity. When the relative velocity, u, of the two observers is much less than the speed of light *in vacuo*, c, these equations reduce to the classical velocity addition rules —equations (1.13), (1.14) and (1.15).

In the case where the 'object', whose velocity is measured by the observer S′ is a light ray, i.e., $v'_{x'} = c$, the velocity that the observer S will measure is

$$v_x = \frac{v'_{x'} + u}{1 + {uv'_{x'}}/{c^2}} = \frac{c + u}{1 + {uc}/{c^2}} = c$$

This result confirms that these velocity addition rules conform to the principle of the constancy of the speed of light.

Similarly, the rules conform to the requirement that no object or signal can move at a speed greater than c. Thus, whereas in classical mechanics, a velocity of $0.9c + 0.5c = 1.4c$ is needed to overtake an object moving at $0.9c$ with a relative velocity of $0.5c$, according to the relativistic equation (1.88), if $u = 0.9c$ and $v'_{x'} = 0.5c$, the required velocity is:

$$v_x = \frac{v'_{x'} + u}{1 + {uv'_{x'}}/{c^2}} = \frac{0.5c + 0.9c}{1 + \left({0.9c \cdot 0.5c}/{c^2}\right)} = 0.9655c$$

i.e., the required velocity is less than c.

1.5.2 The Kinetic Energy of a Particle in Special Relativity

The kinetic energy, K, of a particle moving parallel to the X-axis is equal to the work done in accelerating the particle from rest by the force, F_x, acting on the particle parallel to the X-axis:

$$K = \int_0^x F_x \cdot dx \tag{1.91}$$

Substituting the relevant relativistic definition of force, $F_x = \frac{dp_x}{dt} = \frac{m_0 d(\gamma u)}{dt}$, from equation (1.56) gives:

$0.9c + 0.5c = 1.4c$ is needed to overtake an object moving at $0.9c$ with a relative velocity of $0.5c$, according to the relativistic equation (1.88), if $u = 0.9c$ and $v'_{x'} = 0.5c$, the required velocity is:

$$v_x = \frac{v'_{x'} + u}{1 + {uv'_{x'}}/{c^2}} = \frac{0.5c + 0.9c}{1 + \left({0.9c \cdot 0.5c}/{c^2}\right)} = 0.9655c$$

i.e., the required velocity is less than c.

1.5.2 The Kinetic Energy of a Particle in Special Relativity

The kinetic energy, K, of a particle moving parallel to the X-axis is equal to the work done in accelerating the particle from rest by the force, F_x, acting on the particle parallel to the X-axis:

$$K = \int_0^x F_x \cdot dx \tag{1.91}$$

Substituting the relevant relativistic definition of force, $F_x = \frac{dp_x}{dt} = \frac{m_0 d(\gamma u)}{dt}$, from equation (1.56) gives:

$$K = m_0 \int_0^x \frac{d(\gamma u)}{dt} dx \tag{1.92}$$

The accelerated particle acquires a velocity $u = {dx}/{dt}$ parallel to the X-axis. In terms of this velocity, the particle's kinetic energy is given by:

$$K = m_0 \int_0^u u d(\gamma u) \tag{1.93}$$

Applying the rule, $\int x dy = xy - \int y dx$, for integrating by parts, we obtain:

$$K = m_0 \gamma u^2 - m_0 \int_0^u \gamma u du \tag{1.94}$$

Bcause u is always less than c, we can use the following substitution in order to evaluate the integral:

$$u = c \sin \alpha \tag{1.95}$$

where c is the speed of light *in vacuo*. This, in turn, produces the following additional relationships:

$$du = c \cos \alpha . d\alpha \qquad \gamma = {1}/{\cos \alpha} \tag{1.96}$$

Substituting these relationships in equation (1.80) gives:

$$K = \frac{m_0 c^2 \sin^2 \alpha}{\cos \alpha} - m_0 \int_0^{\alpha} \frac{c \sin \alpha (-c \cos \alpha)}{\cos \alpha} \cdot d\alpha \tag{1.97}$$

$$= \frac{m_0 c^2 \sin^2 \alpha}{\cos \alpha} + m_0 c^2 \int_0^{\alpha} \sin \alpha \cdot d\alpha$$

Evaluating the integral gives:

$$K = \frac{m_0 c^2 \sin^2 \alpha}{\cos \alpha} + m_0 c^2 \left[\cos \alpha\right]_0^{\alpha} \tag{1.98}$$

$$= \frac{m_0 c^2 \sin^2 \alpha}{\cos \alpha} + m_0 c^2 (\cos \alpha - 1)$$

$$= m_0 c^2 \left(\frac{\sin^2 \alpha}{\cos \alpha} + \cos \alpha - 1 \right)$$

Simplifying this result gives:

$$K = m_0 c^2 \left(\frac{\sin^2 \alpha + \cos^2 \alpha - \cos \alpha}{\cos \alpha} \right) \tag{1.99}$$

$$= m_0 c^2 \left(\frac{1}{\cos \alpha} - 1 \right)$$

Returning to the original format by making the substitution $1/\cos\alpha = \gamma$ gives the following relativistic expressions for the particle's kinetic energy:

$$K = m_0c^2(\gamma - 1) \tag{1.100}$$
$$= c^2(\gamma m_0 - m_0)$$
$$= c^2(m - m_0)$$

1.5.3 The Total Energy of a Particle

The total energy, E, of a particle can be expressed in terms of its rest mass, m_0, and its linear momentum, p, through a synthesis of the following four fundamental relationships:

$$\gamma = \sqrt{\frac{1}{1 - u^2/c^2}} \tag{1.29}$$

$$m = \gamma m_0 \tag{1.55}$$

$$E = mc^2 \tag{1.68}$$

$$p = mu \tag{1.54}$$

Algebraic elimination of the particle's relativistic mass, m, its velocity, u, and the factor γ from these equations produces the relationship

$$E^2 = m_0^2 c^4 + p^2 c^2 \tag{1.101}$$

1.5.4 The Transformation of Force

The components of a force acting on a particle relative to the inertial observers S and S′ are:

$$F_x = \frac{dp_x}{dt} \qquad F_y = \frac{dp_y}{dt} \qquad F_z = \frac{dp_z}{dt} \tag{1.102}$$

$$F'_{x'} = \frac{dp'_{x'}}{dt'} \qquad F'_{y'} = \frac{dp'_{y'}}{dt'} \qquad F'_{z'} = \frac{dp'_{z'}}{dt'}$$

In the special case in which the particle is momentarily at rest relative to the observer S, the force he measures will be the *proper force*.

In general, the S components of the force can be converted in those of S′, or *vice versa*, by making the appropriate substitutions in the above definitions. For example, to convert the S′ definitions into those of S, the appropriate expressions for the infinitesimals $dp'_{x'}$, $dp'_{y'}$, $dp'_{z'}$ and dt' in terms of dp_x, dp_y, dp_z and dt are substituted in the definitions of the S components. These expressions are obtained by taking the derivatives of the following momentum and time transformation equations.

$$p'_{x'} = \gamma\left(p_x - {uE}/{c^2}\right) \qquad p'_{y'} = p_y \qquad p'_{z'} = p_z \qquad t' = \gamma\left(t - {ux}/{c^2}\right)$$

In the special case where the particle is momentarily at rest relative to the observer S, i.e., when $dx = 0$ and $dE = 0$, the relevant substitutions are:

$$dp'_{x'} = \gamma dp_x \qquad dp'_{y'} = dp_y \qquad dp'_{z'} = dp_z \qquad dt' = \gamma dt \tag{1.103}$$

Making these substitutions in the definitions of the S′ components gives

$$\begin{aligned} F'_{x'} &= \frac{dp'_{x'}}{dt'} & F'_{y'} &= \frac{dp'_{y'}}{dt'} & F'_{z'} &= \frac{dp'_{z'}}{dt'} \\ &= \frac{\gamma dp_x}{\gamma dt} & &= \frac{dp_y}{\gamma dt} & &= \frac{dp_z}{\gamma dt} \\ &= F_x & &= {F_y}/{\gamma} & &= {F_z}/{\gamma} \end{aligned} \tag{1.104}$$

Part Two Quantum Theory

Despite its revolutionary notions concerning the nature of space and time, Einstein's special theory of relativity did not require any changes in Maxwell's electromagnetic theory.

Even the realization that there is no aether, astounding as it was at the time, could be taken in its stride.

Notwithstanding Maxwell's own conviction that an aether did exist, it was not a crucial element of his theory.

However, the nonexistence of the aether did raise profound questions about the actual physical nature of electromagnetic waves.

For if there is no aether, what medium carries the energy transported by an electromagnetic wave?

How can empty space carry energy?

How does sunlight reach us across 150,000,000 km of empty space?

And what reality can we ascribe to a wave that propagates through empty space?

There was nothing in special or general relativity that could answer these questions.

The existence or otherwise of the aether was not the only troubling issue faced by electromagnetic theory towards the end of the 19th century. Systematic quantitative studies of the emission and absorption of light and heat carried out at the time had uncovered phenomena that were totally inexplicable in the framework of Maxwell's theories.

There appeared to be a fundamental flaw in the wave theory of light.

The new theory - *Quantum Theory* - that was developed in the first decades of the 20th century to explain these phenomena, provided an answer to the question of the physical nature of electromagnetic waves.

Subsequently, this same theory furnished the conceptual framework for **a** new understanding of the structure of matter and of its chemical and physical properties.

However, like the theory of relativity before it, quantum theory also requires fundamental changes in our intuitive 'common-sense' grasp of reality.

This time, it is not the equivalence or otherwise of our observations that is questioned but their very nature and that of their underlying reality.

What do we actually mean when we say: "the mass of the electron is 9.31×10^{-31}kg"?

That a particle's mass is a property that adheres to it; something it actually has?

Or is it the result of a measurement, that exists only through the observer's consciousness?

In general, does a measurement reveal the particular result it does because that is the value the quantity actually had at the moment it was measured?

Or did the quantity only exist *potentially* before being brought into being by the act of measurement?

And what if the potentiality allows for more than one possible outcome of a measurement?

How are the various possibilities disentangled when a measurement is actually made and just a single outcome is obtained?

We shall be concerned with these and many similar esoteric questions during our investigation of Quantum Theory.

Part 2.1 The Quantum Hypothesis

2.1.1 Radiators and Radiation

The energy emitted and absorbed by material objects in the form of electromagnetic waves is generally termed *electromagnetic radiation* or simply *radiation.*

Radiation is emitted at the expense of the energy of the radiating object; absorbed radiation adds to the energy of the absorbing object.

Any specific distribution of electromagnetic radiation, such as the colors of the rainbow, is termed a *spectrum.*

If the distribution comprises a continuous region of frequencies or wavelengths, it is a *continuous spectrum;* if it comprises a series or group of discrete frequencies, it is a *line spectrum.*

The distribution of the electromagnetic radiation emitted by a body that acts as a source of radiation is its *emission spectrum* (Fig 2.la).

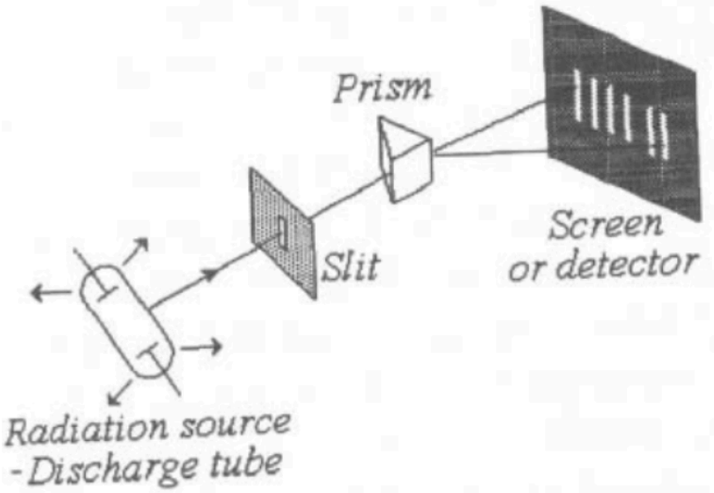


Fig2.1a Viewing an emission spectrum in the visible range. The light emitted from the source is dispersed into its components by a prism or diffraction grating. These components (colors) illuminate the screen or detector and constitute the source's emission spectrum.

The rainbow is the sun's emission spectrum in the visible region.

The distribution of the electromagnetic radiation transmitted by an absorbing medium placed in the path of radiation that exhibits a continuous spectrum, is called an *absorption spectrum* (Fig 2.lb).

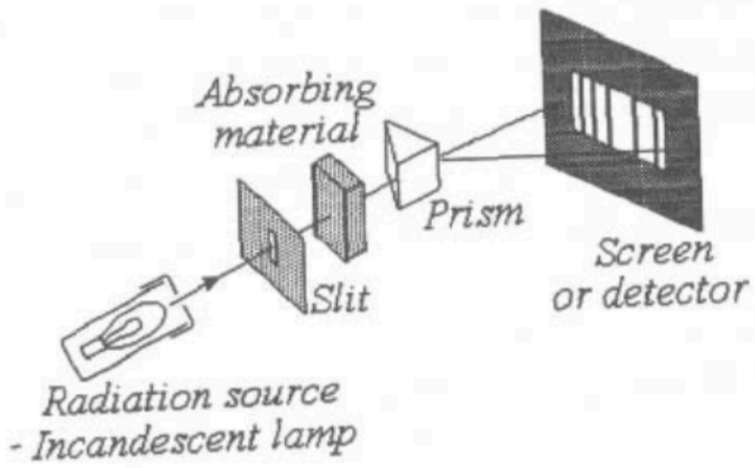


Fig 2.lb Viewing an absorption spectrum in the visible range. Light, such as that emitted by an incandescent lamp which exhibits a continuum of frequencies, is passed through a sample of the material whose absorption spectrum is being investigated. The frequencies absorbed by the sample appear as dark lines or bands on the continuous spectrum projected onto the screen.

When viewing a rainbow through a colored filter, certain colors are seen to be missing; they are the ones absorbed by the filter.

The spectrum of the transmitted light is the filter's absorption spectrum.

In general, the electromagnetic radiation emitted and absorbed by a body depends on two factors:

1) its temperature;

2) the characteristics of the body and of the material from which it is made.

Quantum theory evolved from the attempts to find the correct functional relationship between these factors and the frequencies and intensities of the electromagnetic radiation that material bodies emit and absorb.

2.1.2 Thermal Radiation

The electromagnetic radiation that bodies emit and absorb by virtue of their temperature is called *thermal* or *heat radiation*.

The spectrum of heat radiation is a continuous one that comprises all the wavelengths (or frequencies) in a specific region of the electromagnetic spectrum.

A plot of the intensity of the heat radiation emitted by an object as a function of the radiation's wavelength -the *intensity spectrum* - resembles a statistical distribution (Fig 2.2).

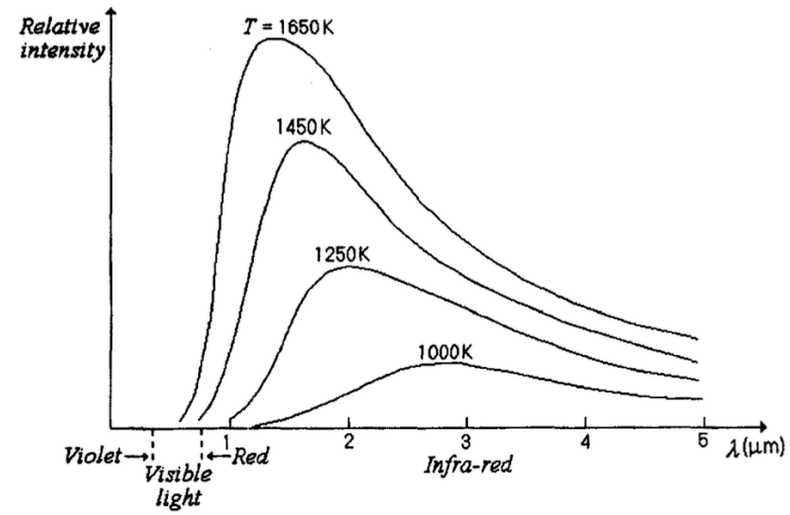


Fig 2.2 The intensity spectrum of the heat radiation emitted by an object at different temperatures. **As** the temperature rises the spectrum includes shorter wavelengths. Likewise, the wavelength of maximum intensity - the wavelength of the dominant color of the radiation decreases with rising temperature.

The wavelengths of the heat radiation emitted from a body at a temperature below 1000K are in the invisible infra-red region of the electromagnetic spectrum: $\lambda > 1.2\mu$m.

As the temperature rises, radiation of lower wavelength appears in the spectrum.

At about 1200K the spectrum includes wavelengths in the visible red, $\lambda_{Red} \approx 0.7\mu$m.

Raising the temperature further, introduces still shorter wavelengths and the color of a hot glowing object changes accordingly, first to yellow and finally to white.

The temperature of the element in an incandescent lamp is almost 3000K.

In 1792, Prevost' proposed that all bodies are in fact emitting heat all the time, but that they do so at a slower rate at low temperatures than at higher ones.

Hence, an object placed in surroundings whose temperature is higher than its own is irradiated by a greater flux of heat energy than it emits.

The object absorbs the difference and its temperature rises accordingly until the flow of heat energy it emits exactly balances that it absorbs from its surroundings.

From that point on, the temperature of the object remains constant, equal to that of its surroundings.

The object is in thermal equilibrium, absorbing heat from its surroundings at the same rate as it emits.

Heat radiation is formed inside a body by the random thermal motions of its atoms and molecules.

On its way to the surface, this radiation is repeatedly absorbed and re-emitted by the body's particles.

In this process, the original character of the radiation is totally obliterated.

Hence, irrespective of the identity of the object's atoms and molecules, as a result of this smoothing out process, the intensity spectrum of the heat radiation inside every object at a given temperature is the same.

However, not all the heat radiation reaching the object's surface from the inside is emitted, just as not all of that incident on it from outside is absorbed.

The flux of heat radiation actually emitted or absorbed by an object depends on the nature of its surface.

Different kinds of surfaces absorb and emit heat radiation to a different extent.

Black surfaces are good absorbers, whereas polished or white surfaces are not.

From Prevost's *theory of exchanges* it follows that good absorbers must be equally *good* emitters, and poor absorbers must be equally poor emitters.

Were this not so, an object that was a good absorber of heat radiation but only a poor emitter, could not maintain thermal equilibrium with its surroundings.

It would absorb too much heat radiation, emit too little and so spontaneousIy warm up to a temperature above that of its surroundings.

But this would be contrary to our common experience.

After reaching thermal equilibrium, objects do not spontaneously raise their temperature above that of their surroundings; nor do they spontaneously lower it.

Evidently, at thermal equilibrium, all objects emit heat radiation that is equivalent in its quality and quantity to that which they absorb

2.1.3 Black-body Radiation

Common materials and objects do not absorb all the radiation incident upon them; they are not perfect absorbers of radiation.

Nevertheless, we can imagine a perfect absorber, an ideal body which does absorb all the electromagnetic radiation that strikes it, whatever its wavelength or intensity.

Because such a body would appear black in whatever light it is illuminated, it is called a *black-body*.

Since, by definition, a black-body is a perfect absorber, it must also be a perfect emitter, i.e., it must be able to emit radiation of every wavelength at any intensity.

The heat radiation emitted by a black-body is called *black-body radiation*.

No substance is a perfect black-body, though soot, which absorbs some 95% of the visible and infrared radiation incident upon it, closely approximates to one in this range.

However, by making a small hole in the wall of a hollow object we can construct a device, which to all intents and purposes, absorbs all the radiation incident upon it.

The device works somewhat like a fly-trap.

Any radiation coming upon the hole from outside will pass through it, enter the cavity and be trapped inside (Fig 2.3).

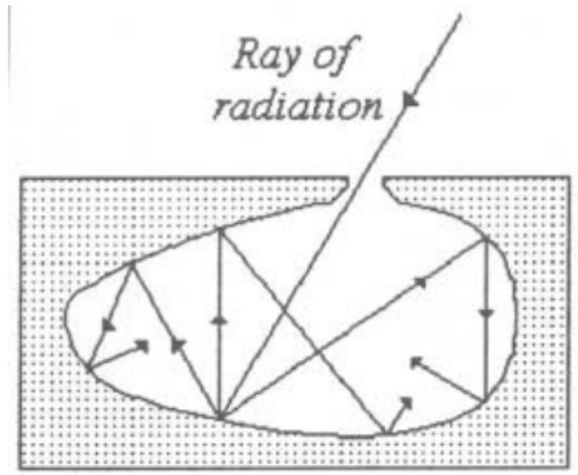


Fig 2.3 A hollow object with a small aperture in one of its walls. Any radiation that enters through the hole will be trapped inside.

The hole is thus a perfect absorber of radiation; it is a black-body.

By definition, it will also be a perfect emitter of radiation.

Any radiation generated in the cavity that happens upon the hole will escape without hindrance, irrespective of its wavelength or intensity,

Suppose, that a hollow object with a small aperture in one of its walls is kept at a constant temperature.

Every point on the inside surface of the cavity will be in thermal equilibrium with all the other points on the surface and heat radiation of the same quality and quantity will be emitted and absorbed by each point, irrespective of the material from which the inside surface of the cavity is made.

The cavity will be filled by electromagnetic radiation that comprises all the wavelengths of the heat radiation characteristic of the object's temperature, each at its appropriate intensity.

Any of this *cavity radiation* that happens upon the aperture from inside will escape through it unhindered.

Viewed from outside, this representative sample of the cavity radiation will be the black-body radiation characteristic of the particular temperature.

Experimental investigations of black-body radiation conducted towards the end of the 19th revealed the following three important features:

1. **The intensity of the heat radiation emitted from a black-body is greater than that emitted by any other body at the same temperature (Fig 2.4).**
2. **The energy, E, radiated each second from each unit area of the surface of a black-body is directly proportional to the fourth power of the body's absolute temperature, T:**

$$E = \sigma T^4 \qquad (2.1)$$

where the proportionality constant $\sigma = 5.672 \cdot 10^{-8} \mathrm{W \cdot m^{-2} \cdot K^{-4}}$. This rule is known as the *Stefan*[3]*-Boltzmann*[4] *Law*.

3. **The absolute temperature, T, of a black-body is inversely proportional to the wavelength *in vacuo*, λ_{max}, of the radiation it emits with the greatest intensity (Fig 2.5):**

$$T\lambda_{max} = 0.29 \cdot 10^{-2}\,\mathrm{m \cdot K} \qquad (2.2)$$

This rule is known as *Wien's Law*.

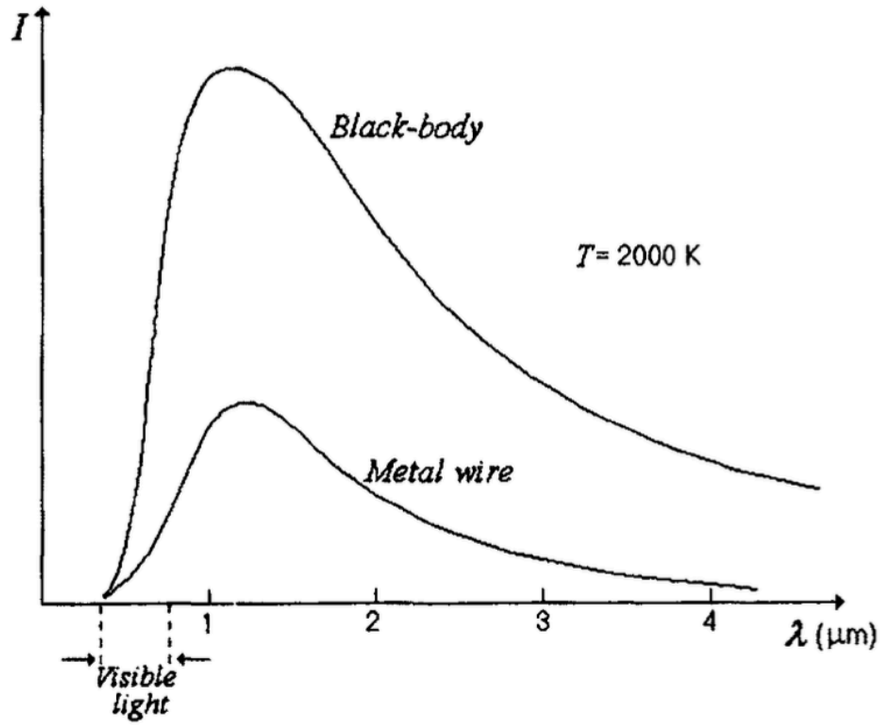


Fig 2.4 The intensity, I, of the heat radiation emitted by a black-body and the filament in an incandescent lamp at a temperature of 2000K as a function of the wavelength, λ, of the radiation.

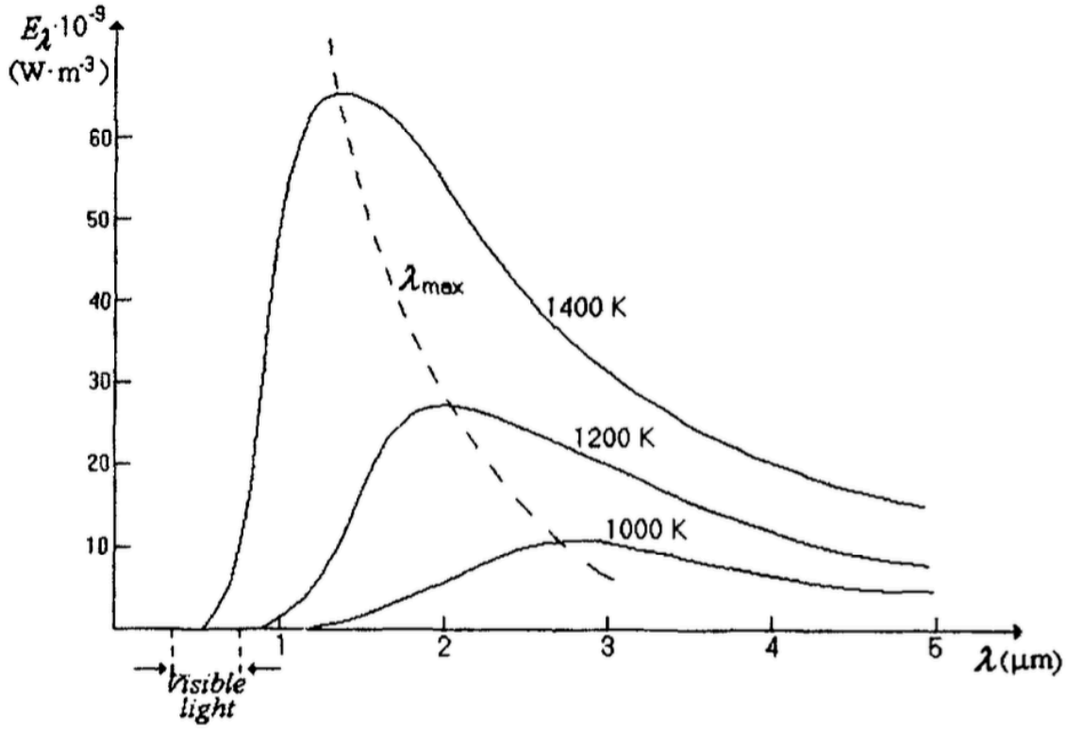


Fig **2.5** The distribution of the heat energy radiated by a black-body between the various wavelengths at a number of temperatures E_λ is defined such that $E_\lambda . \Delta\lambda$, is the energy radiated each second at the wavelengths between λ and $(\lambda + \Delta\lambda)$ from a unit area of the body's surface. The area under each curve gives the total energy radiated each second from each unit area at the particular temperature.

The best physical evidence for the ‘big-bang’ hypothesis in cosmology is the radio frequency radiation, that appears to be black-body radiation corresponding to a temperature of 3K, discovered coming from outer space.

The intensity of the radiation is almost completely uniform in all directions, which suggests that the universe itself is its source.

According to the hypothesis, the universe arose out of a point singularity about 15 to 20 billion years ago.

At the universe’s beginning, when its temperature was extremely high, the radiation and elementary particles were in a kind of dynamic equilibrium, in which particles were unceasingly changing into radiation and the radiation back into particles.

Under these conditions, neutral atoms and molecules could not form.

With the passage of time, the universe expanded and cooled and when its temperature had dropped t0 3000K, normal matter was first formed.

At this stage the synergetic link between radiation and matter was broken and the radiation, which was black-body radiation corresponding to a temperature of 3000K, began to propagate through the expanses of space.

It is hypothesized, that the ubiquitous 3K black-body radiation is the residue of the radiation released when matter and radiation divorced, the temperature of the radiation having fallen a thousand-fold since then as a result of the subsequent further expansion and cooling of the universe.

2.1.4 Difficulties in the Classical Theory of Radiation

Finding an explanation for the intensity spectrum of the thermal radiation emitted by a black-body was one of the most pressing unsolved problems of physics in the last decade of the 19th century.

All the attempts to derive a mathematical function that correctly predicts the intensity spectra in Fig 2.5 from classical postulates failed.

Without going into all the details, we can explain the basic reasons for this failure.

According to Newtonian and Maxwellian physics, the average energy, $\bar{E}_{Oscillator}$ of an oscillator that is in thermal equilibrium with its surroundings depends only on its absolute temperature, ***T:***

$$\bar{E}_{Oscillator} = kT \tag{2.3}$$

The constant of proportionality in this equation, $k = 1.38 \times 10^{-23} JK^{-1}$, is called *Boltzmann*'s *Constant*.

This relationship is derived from the *Principle of the Equipartition of Energy* according to which all the oscillators in a given system will possess the same average energy at equilibrium, irrespective of their frequencies.

For example, in a system of 100 oscillators, 20 with a natural frequency of l000Hz and 80 with a natural frequency of 5000Hz, on average 20% of the system's energy should be associated with the low frequency oscillators and 80% with the higher frequency ones.

The electric oscillators that generate the heat radiation in a material object are the particles - the atoms and molecules - of which it is composed.

The frequencies of the heat radiation emitted by the object are those at which these particles can vibrate.

It follows that the more frequencies there are in a particular range, the greater will be the intensity of the radiation in that range.

If, for example, a system's oscillators can vibrate at 10 different frequencies in the range between 500Hz and 1500Hz, but at 100 different frequencies in the range between 7500Hz and 8500Hz, then the energy emitted in the latter range will be ten times greater than that emitted in the former.

In general, the more permitted frequencies there are in a given region of the spectrum, the greater will be the emission intensity in that region.

Thus, the problem of explaining the intensity spectrum of black-body radiation, reduces to determining the number of frequencies allowed to the black-body's electrical oscillators in each region of the spectrum**.**

It can be shown that the electrical oscillators in a system such as the cavity of a constant-temperature hollow body possess many more high frequency vibrational modes than low frequency ones.

Consequently, if the energy of a cavity that is in thermal equilibrium is allocated on average equally between all the allowed vibrational modes, as required by the equipartition principle, most of this energy will be associated with the higher frequencies.

Thus, cavity radiation of any temperature should be predominantly very high frequency radiation, i.e., it should comprise just a little infra-red (low frequency radiation), lots of ultra-violet (high frequency radiation) and even more X-rays (very high frequency radiation).

However, this is contrary to experience.

Black-body radiation always exhibits its maximum intensity near the middle of its range of frequencies and not at the high frequency end.

In practice, the highest frequencies are not operative; the system's energy is not in fact equipartitioned between all its frequencies (Fig 2.6).

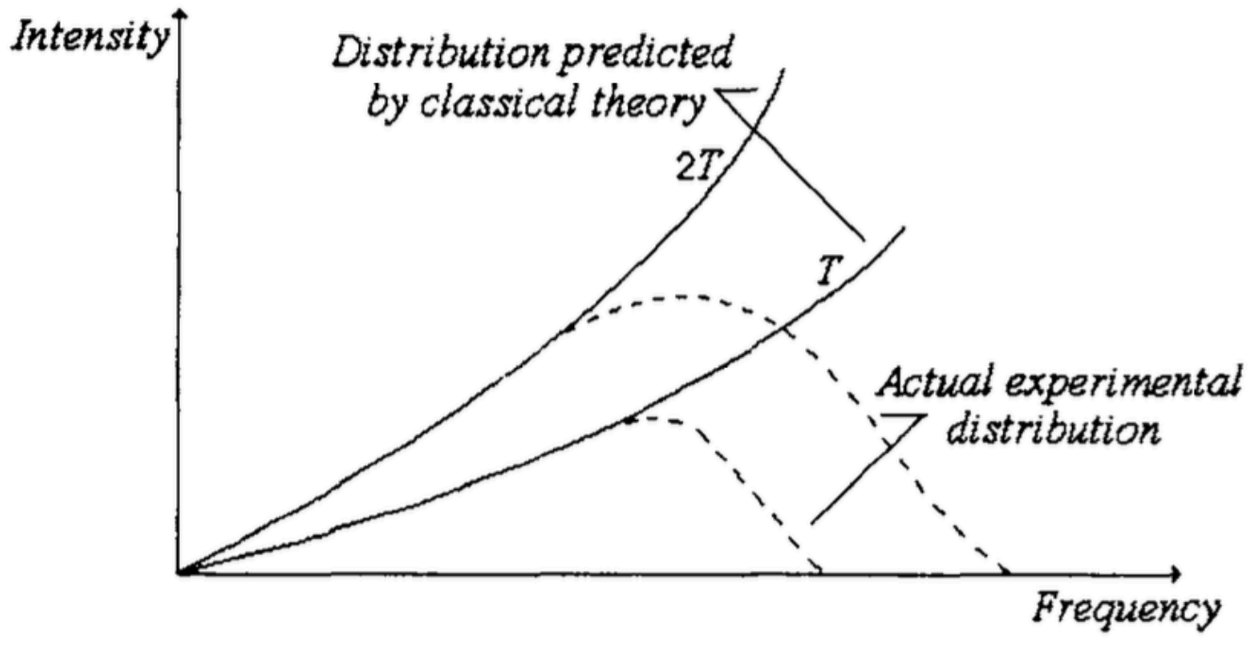


Fig 2.6 The distribution of the intensity of black-body radiation as a function of its frequency at two temperatures, T and 2T. The distribution predicted by classical physics - the continuous line, coincides with that found experimentally - the broken line, only at low frequencies. The area under each curve represents the total energy of the black-body radiation. Whereas the areas under the experimental curves are finite, those predicted by classical physics are potentially infinite, which is obviously an impossibility.

When first presented, this construal of heat radiation was termed the ultra-violet catastrophe, since it implied that every physical system should spontaneously explode in a burst of high frequency radiation.

2.1.5 Planck's Quantum Hypothesis

By the end of the 19th century, the atomic nature of matter had been well established.

It was recognized that matter could only exist in integer multiples of a natural unit - one hundred atoms, a mole of atoms, etc. - and could only be allocated in discrete amounts.

However, as regards energy, it was still believed that it could be absorbed or emitted continuously in any arbitrary amount.

This was a necessary corollary of the equipartition principle, which required the equal and unhindered distribution of a system's energy between all its vibrational modes irrespective of their frequencies.

In 1900, by abandoning this tenet, Planck' first succeeded in deriving a function that correctly described black-body radiation.

Suggesting that the energy of an oscillator is not "a continuous, infinitely divisible quantity, but a discrete quantity composed of an integral number of finite equal parts", Planck showed that Wien's law required that the energy of each of these finite equal parts be directly proportional to the oscillator's frequency, f.

On this basis, he argued that the oscillator's energy can only take the specific values

$$E_{Oscillator} = 0, hf, 2hf, 3hf, \ldots \tag{2.5}$$

all others being forbidden. The constant of proportionality, $h = 6.63x10^{-34}$ J s , is the universal constant now called Planck's Constant.

It follows, that an oscillator's energy can only change by certain discrete amounts, i.e., that it can emit or absorb energy only in small 'packets' -what are now called quanta- the energy of such a packet or quantum depending on the oscillator's frequency, such that

$$E_{Quantum} = hf \tag{2.6}$$

It was on the basis of this hypothesis -the *Quantum Hypothesis*- that Planck succeeded to derive a mathematical function which correctly predicted the intensity spectrum of black-body radiation. In doing so, Planck showed that the average energy of an oscillator in thermal equilibrium at an absolute temperature T is correctly given by the equation

$$\bar{E}_{Oscillator} = \frac{hf}{e^x - 1} \tag{2.7}$$

where the factor x is given by the equation

$$x = \frac{hf}{kT} \tag{2.8}$$

and not by the classical relationship $\bar{E}_{Oscillator} = kT$. Thus, energy is not allocated equally between a system's oscillators; each oscillator's average energy depends not only on its temperature but also on its frequency.

We can now explain, qualitatively, why the ultra-violet catastrophe does not happen.

At normal temperatures, the energy, h*f* of the quanta associated with radiation of high frequency is much greater than the average kinetic energy, kT, of the atoms comprising the source of the radiation, i.e., $hf >> kT$.

Under these conditions, the probability of any atom (oscillator) acquiring sufficient energy to create such a high energy quantum is negligible.

Thus, the oscillators' high frequencies are not operative and so the heat radiation emitted does not contain high frequency radiation.

At low frequencies, the energy of the quanta is much smaller than the average kinetic energy of the atoms, i.e., $hf << kT$.

In this case, the factor x in equation (2.7) will be correspondingly small and to a close approximation we can say that

$$e^x = 1 + x = 1 + {hf}/{kT} \tag{2.9}$$

Substituting this in equation (2.7) gives

$$\overline{E}_{Oscillator} = \frac{hf}{\left(1 + {hf}/{kT}\right) - 1} = kT \tag{2.10}$$

which is the classical equation (2.10) for the average energy of an oscillator at equilibrium.

The energy quanta associated with changes in the energies of low frequency oscillators are so minute that changes in the oscillators' energies appear to be continuous.

This is why the classical approach correctly predicts the actual experimental distribution of intensities at low frequencies.

Despite the success of the quantum hypothesis in explaining black-body radiation, Planck himself continued to believe that the radiation itself was a pure wave phenomenon.

It was difficult to accept the idea that a natural process was discontinuous.

This idea stood in total contradiction to one of the most venerable postulates of western thinking - *natura non facit saltum* - nature does not make jumps.

Only after the quantum hypothesis had repeatedly demonstrated its competence in explaining different physical phenomena did it gain acceptance as a fundamental doctrine of physics.

A guideline for the coexistence of the classical physics of Newton and Maxwell with the new quantum physics was first formulated by Niels Bohr.

Known as the Correspondence Principle, it states that:

> **The results predicted by quantum physics must be identical to those predicted by classical physics in those situations where classical physics corresponds to the experimental facts.**

This principle reconfirms the rule that in the physical sciences it is the experimental facts that are the final arbiter.

Whatever else changes, a theory remains useful and valid so long as it satisfies this requirement and correctly predicts the facts .

2.1.6 Atomic Spectra

In contrast with the typical continuous spectrum of thermal radiation, the spectrum of the light emitted from a low pressure discharge tube containing an atomic gas such as atomic hydrogen, helium or sodium vapor is a line spectrum, each line corresponding to a specific frequency or wavelength (Fig 2.7)**.**

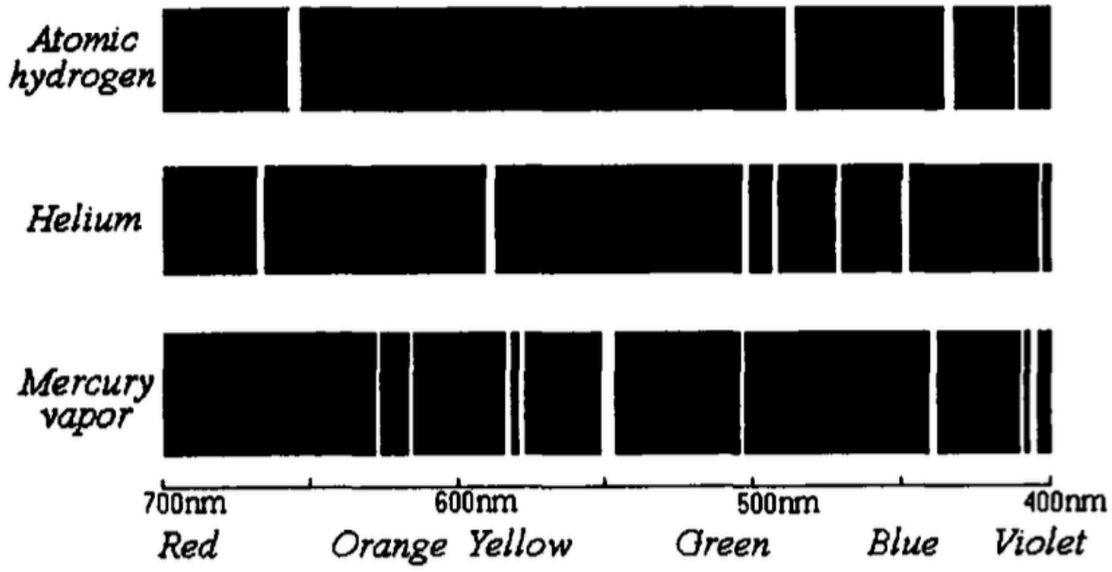


Fig *2.7* Emission spectra of a number of atomic gases. The light emitted by the excited gases comprises discrete frequencies which appear as lines in the spectrum.

Atomic emission spectra can also be observed by analyzing the light emitted from the gas when heated to a high temperature.

The absorption spectra of atomic gases are also line spectra.

When white light is passed through an atomic gas, some of the components of the light are absorbed.

The absorbed components are absent in the transmitted light.

Their absence shows up in the spectrum of the emergent light as dark lines on the illuminated colored background of its continuous spectrum (Fig 2.8).

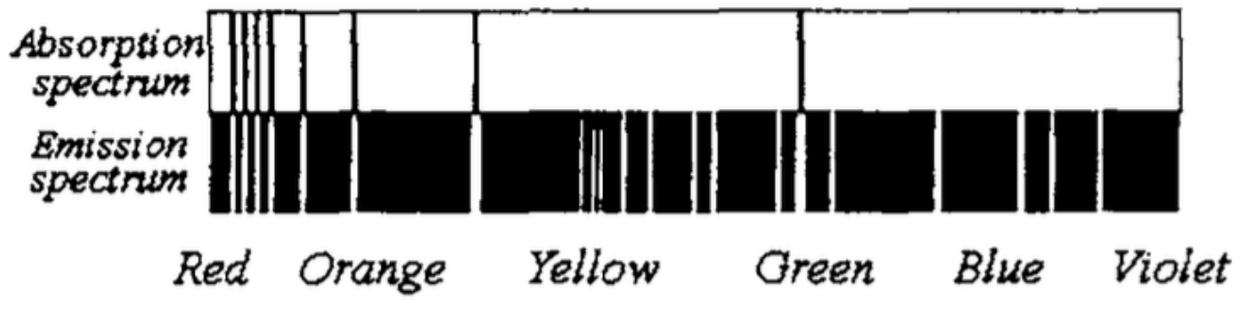


Fig 2.8 The absorption and emission spectra of sodium vapor. Note that there are less lines in the absorption spectrum than in the emission spectrum and that all the absorption frequencies correspond to those of emission lines. Note also the differences in the intensities of the various lines.

The frequencies of the absorption lines generally correspond to those of lines in the emission spectrum.

However, there are generally less lines in the absorption spectrum of an element than in its emission spectrum.

Light absorption is an example of resonance; the atoms in the gas absorb energy from the incident white light at frequencies that match one of their own natural frequencies.

In the middle of the 19th century, Kirchoff demonstrated that each chemical element exhibits a distinctive emission and absorption spectrum; an element's spectra are its unique finger-prints.

Thus, chemical elements can be identified by means of their spectra.

On the basis of this insight, Kirchoff was able to elucidate a phenomenon first observed some fifty years earlier by Fraunhofer in his investigations of the spectrum of sunlight.

Since sunlight is the heat radiation characteristic of a temperature of about 6000K, its spectrum should be continuous.

However, when Fraunhofer examined the spectrum of sunlight closely, he found it crossed by some 700 dark lines.

Kirchoff explained that these are absorption lines, introduced into the sunlight as it passes though the cool outer layers of the sun's chromosphere

The presence of some seventy chemical elements in the sun's chromosphere has been established by the study of the spectrum of sunlight.

The element helium was observed spectroscopically in the sun's chromosphere some thirty years before it was isolated on earth.

All our information about the composition of the stars and the galaxies is drawn from the spectra of starlight.

And it is from shifts in the wavelengths of the spectral lines in starlight which are attributed to a Doppler effect, that most of our knowledge and ideas about the origin and expansion of the universe have been gleaned.

Since the emission of light was thought to result from some sort of vibration, harmonic relationships like those found in vibrating mechanical systems were sought between the frequencies of the spectral lines.

Such harmonic series were not discovered.

However, early in the twentieth century, after more data had been collected, Ritz noticed that in many instances the sum of the frequencies of two lines in an atom's spectrum gives the frequency of a third line and that the sum of the frequencies of these three lines gives the frequency of a fourth line, and so on.

This is called the *Ritz Combination Rule* (Fig 2.9).

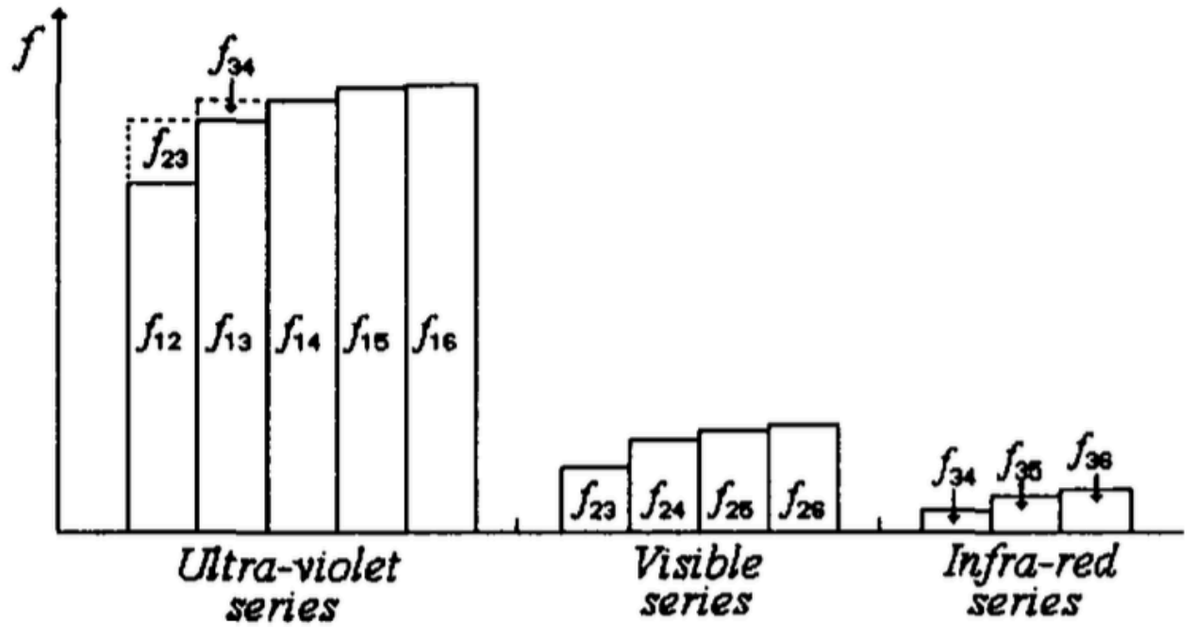


Fig 2.9 Ritz's combination rule. The lines in an imaginary atom's spectrum**.** The spectral lines are identified by the two subscript numbers of each line's frequency. The first number indicates the series or group to which the line belongs. The diagram shows the following relationships: $f_{12} + f_{23} = f_{13}$; $f_{12} + f_{23} + f_{34} = f_{14}$.

Ritz's empirical rule was first given a theoretical foundation in 1913**,** when Bohr**,** in the framework of his researches on the structure of the atom, employed Planck's quantum hypothesis to explain the origin of atomic line spectra.

According to the quantum hypothesis, an electric oscillator such as an atom absorbs and emits radiation in discrete quanta, each of energy hf where f is the frequency of the radiation.

The apparently continuous absorption and emission of radiation by a body is in fact the smoothed out result of a myriad of discrete events, of countless quanta being absorbed or emitted by innumerable atomic oscillators.

Bohr postulated that between these discrete events, when the atoms are not actually emitting or absorbing a quantum of energy, they are in non-radiating stationary states, each of which is characterized by a specific amount of energy, E_1, E_2, E_3,…,called its *energy level* (Fig 2.10).

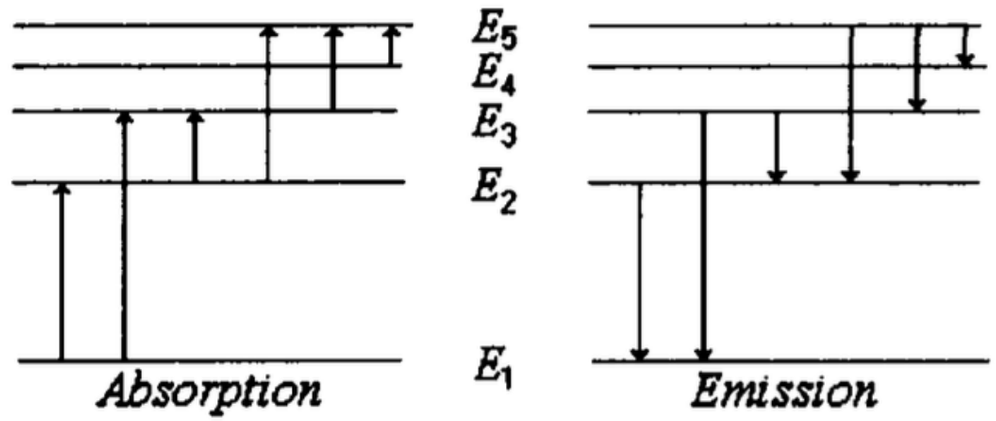


Fig 2.10 A hypothetical system of stationary states showing the transitions between the energy levels that accompany the absorption or emission of energy by the system. The system rises to a higher level when it absorbs energy and drops to a lower one when it emits energy.

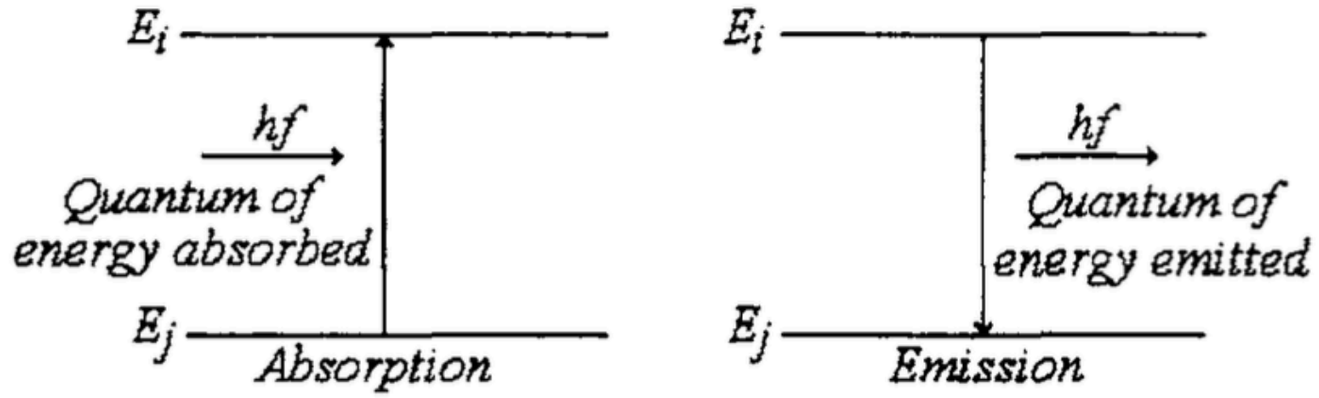
The lowest energy level in a system is its *ground state* and is usually designated E_1.

The system's higher energy levels are called *excited states*.

In passing from one stationary state to another, atoms (oscillators) emit or absorb a quantum of radiation.

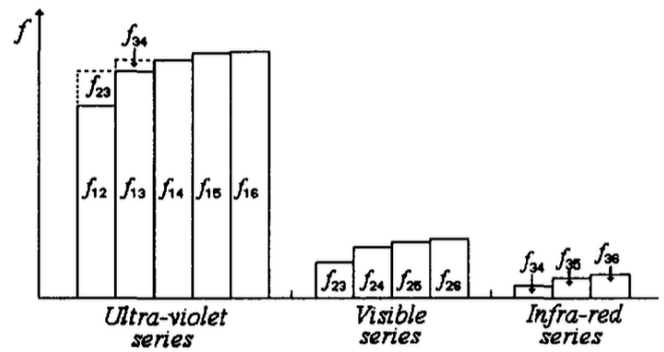
The difference, $E_i - E_j$, between the energy level, E_i, of the initial stationary state, and that of the final stationary state, E_f is balanced by the energy, hf_{ij} of the quantum of radiation associated with transitions between the states (Fig 2.11).

Fig 2.11 The absorption and emission of a **quantum** of energy. When absorbing a photon the atom rises ('jumps') from the lower energy level to the higher one. An atom that emits **a** photon 'falls' to a lower energy level.



Bohr posited that the frequency, f_{ij}, of the radiation emitted or absorbed in a transition is given by

$$f_{ij} = \frac{E_i - E_j}{h} \tag{2.11}$$



Equation(2.11) is known as *Bohr's Equation*.

To appreciate how Bohr's postulate accounts for the Ritz combination rule, consider the relationship $f_{12}+f_{23} = f_{13}$ between the frequencies of the spectral lines of the imaginary atom in Fig 2.9.

Energy conservation requires that there be the same energy change whether the transition from the E1 level to the E3 level takes place directly or by stages, i.e.,

$$(E_1 - E_3) = (E_1 - E_2) + (E_2 - E_3) \tag{2.12a}$$

Dividing this expression by h gives

$$\frac{(E_1 - E_3)}{h} = \frac{(E_1 - E_2)}{h} + \frac{(E_2 - E_3)}{h} \tag{2.12b}$$

From Bohr's equation, the frequencies of the spectral lines f_{i3}, f_{i2} and f_{23}, are:

$$f_{13} = {(E_1 - E_3)}/{h} \qquad f_{12} = {(E_1 - E_2)}/{h} \qquad f_{23} = {(E_2 - E_3)}/{h}$$

respectively.

Substituting these values in equation (2.19b) gives:

$$f_{13} = f_{12} + f_{23}$$

Thus, Ritz's combination rule follows from energy conservation and Bohr's postulate.

With the help of some statistical theory, Bohr's postulate also explains why there are generally less lines in the absorption spectrum of an element than in its emission spectrum, i.e., why gaseous atoms do not usually absorb all the frequencies they emit.

Obviously, an atom cannot absorb the quantum of radiation that corresponds to a transition from a particular state to a higher excited state unless it is already occupying the particular state; it cannot rise from E_3 to E_4 if it is not already in E_3.

In a sample of *gaseous* atoms, the ratio between the number of atoms occupying an excited state, N_i, to the number, N_1, occupying the ground state is given by the Maxwell-Boltzmann distribution(from Statistical Physics):

$$N_i/N_1 = e^{-(E_i - E_1)/kT} \qquad (2.13)$$

where k is the Boltzmann constant and T is the sample's absolute temperature

At room temperature (25°C) the value of the factor kT is

$$kT = 1.36 \cdot 10^{-23} \cdot 298 = 4 \cdot 10^{-21}\,\mathrm{J} \tag{2.14}$$

The energy difference between a pair of adjacent atomic energy levels is typically about 2 eV

$$E_i - E_1 \approx 2eV = 3.2 \cdot 10^{-19}\,\mathrm{J} \tag{2.15}$$

Whence, the ratio between the number of atoms occupying the two energy levels is

$$N_i / N_1 = e^{-2 / 2.57 \cdot 10^{-2}} \approx e^{-80} \tag{2.16}$$

It follows that, at room temperature, the population of the excited state is negligible and the probability of the occurrence of a transition that starts from this state is close to zero.

As a result, under normal conditions, the lines corresponding to transitions from excited states to higher states do not usually appear in the atom's absorption spectrum; the only lines that do appear are those corresponding to transitions that start from the ground state.

With rising temperature, the ratio Ni/N_1 increases, and the population, N_i, of the excited state grows accordingly.

Consequently, the probability of transitions starting from this excited state increases, and so more lines begin to appear in the absorption spectrum.

Ritz's empirical rule was also the starting point for the Matrix Mechanics developed by Heisenberg in 1925.

Asserting that in a logically consistent system no entities should be introduced except such as are physically observable, he listed the observed spectral frequencies emitted by an atom in a square array as follows:

$$\begin{array}{llll} f_{11}=0 & f_{12} & f_{13} & \cdots \\ f_{21} & f_{22}=0 & f_{23} & \cdots \\ f_{31} & f_{32} & f_{33}=0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

Each element in the array corresponds to a transition between a pair of energy levels.

Heisenberg went on to suggest that the values of any observable quantity associated with the transitions could be listed in a corresponding square array.

On this view, physical quantities such as a particle's position and momentum are *not* ordinary functions of time, but represent matrices each of whose elements is associated with a transition from one energy level to an another.

Calculations can be performed with these infinite matrices by the usual rules of matrix theory, subject to certain distinctive commutation rules, which have to take the place of the quantum hypothesis.

2.1.7 The Franck-Hertz Experiment

In 1914, convincing experimental evidence of the stationary atomic states postulated by Bohr was obtained by Franck and Hertz in the course of their investigations of collisions between electrons and gaseous mercury atoms.

In order to study these collisions, the electrodes of a glass triode, containing low pressure mercury vapor, were connected to variable potentials, V_1 and V_2, as shown in Fig 2.12.

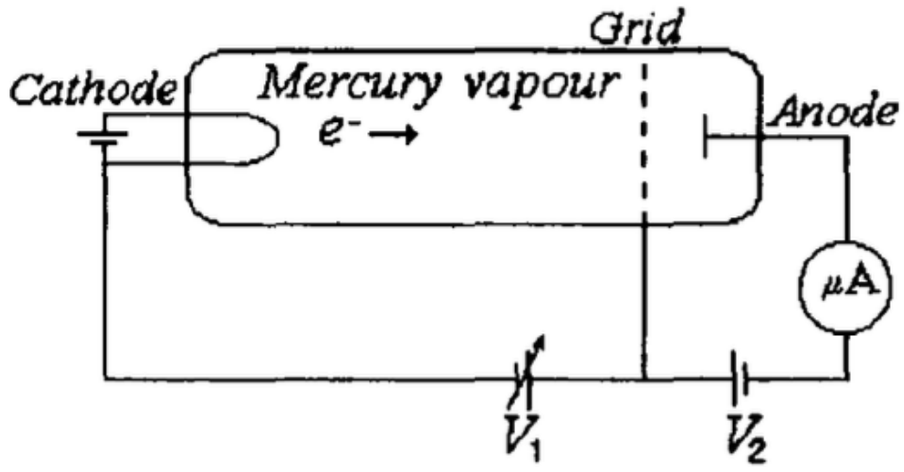


Fig 2.12 Electrons released from the cathode are accelerated by the variable potential V_1 towards the grid. Only those electrons emerging from the grid with an energy greater than eV_2 can reach the anode. Between the cathode and the grid the electrons collide with atoms of the mercury vapor.

Electrons released from the heated cathode C are accelerated towards the grid G by the potential V_1. The anode A was kept at a potential, V_2, lower than the grid.

Consequently, electrons passing through the grid could reach the anode only if their kinetic energy was equal to or greater than eV_2.

The current, I, shown by the microammeter indicated the rate at which electrons arrived at the anode.

In the space between the cathode and the grid, the accelerated electrons collide with atoms of the mercury vapor.

The electrons are very much lighter than the mercury atoms, $m_e \approx m_{Hg}$ /370,000, and so in elastic collisions with the mercury atoms the electrons lose only a minute amount of their energy; the electrons recoil as though they had struck a massive wall.

Under these circumstances, electrons reaching the grid will have sufficient energy to carry on to the anode and a significant anode current, I, will be recorded by the microammeter.

Furthermore, the higher the potential through which the electrons are accelerated, the more of them will reach the anode and the greater will be the anode current.

In their experiment, Franck and Hertz recorded the changes in the anode current, I, as they varied the potential, V_1, between the cathode and the grid.

The results they obtained are shown in Fig 2.13.

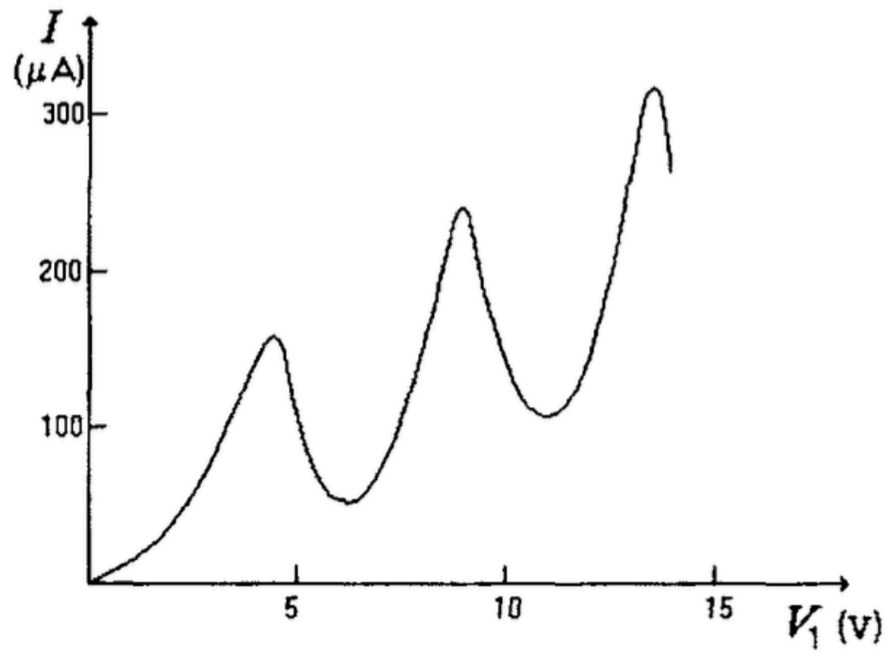


Fig **2.15** Franck-Hertz experiment. The anode current, I, as a function of the potential difference, V_1, between the cathode and the grid.

These show that at first the anode current increases as the potential, V_1, through which the electrons are accelerated is raised; this is consistent with elastic collisions between the electrons and the mercury atoms.

However, at a potential of 4.9V the current falls sharply, almost to zero.

Raising the potential beyond this value restores the current, and it continues to increase as the potential rises until suddenly it drops a second time at a potential of 9.8V.

Raising the potential further, once again restores the current only for it to fall a third time at a potential of about 14.7V.

The sudden changes in the anode current indicate, that at those specific potentials, the number of electrons reaching the anode drops drastically.

It appears as though at these potentials the electrons are 'robbed' of their energy as they pass through the mercury vapor and cannot, therefore, overcome the opposing potential difference, V_2 ,between the grid and the anode.

The simplest explanation for the loss of the electrons' energy is that at these specific potentials, their collisions with the mercury atoms are not elastic.

The inference is, that whereas electrons with other kinetic energies collide elastically with the mercury atoms, electrons with kinetic energies of 4.9eV, 9.8eV or 14.7eV coIlide inelastically with mercury atoms, losing all their energy in the collisions,.

From the point of view of the mercury atoms in the vapor, it appears that, under these conditions, they can only absorb mechanical energy from the electrons in amounts that are an integer multiple of 4.9eV.

These inelastic collisions can be summarized by the symbolic representation:

$$\mathrm{Hg} + \text{fast electron} \rightarrow \mathrm{Hg}^{*} + \text{slow electron}$$

Where Hg* represents a mercury atom in an excited state.

Thus, an electron of energy 4.9eV loses all its energy in a single inelastic collision; an electron of energy 9.8eV in two inelastic collisions; an electron of energy 14.7eV after three inelastic coliisions.

It would appear, that in each of these inelastic collisions, a mercury atom rises from its ground state, E_1, to an excited state, E_2, the difference between their energy levels being 4.9eV.

Unless the energy of the electron is equal to the difference between two energy levels in the mercury atom, or equals an integer multiple of that difference, the atom does not accept any energy from the electron and the collision between them is elastic.

According to Bohr's postulate, equation (2.11), the excited mercury atom, Hg*, can return to its ground state, Hg, by emitting a quantum of radiation of frequency:

$$f = \frac{E_{\mathrm{Hg}^{*}} - E_{\mathrm{Hg}}}{h} = \frac{4.9}{4.14 \cdot 10^{-15}} = 1.18 \cdot 10^{15}\ \mathrm{Hz}$$

Franck and Hertz detected the emission of ultra-violet radiation of this frequency from the glass triode.

This relatively simple experiment demonstrated both the existence of stationary states in the mercury atoms and the validity of Bohr's equation for the relationship between their energy levels and the atom's emission spectra.

In further experiments, it was found that mercury atoms can absorb other specific amounts of energy such as 6.7eV and 8.8eV, indicating the existence of other energy leveIs in the atom.

Part 2.2 The Photoelectric Effect

Though the success of Planck's quantum hypothesis may have established that the absorption and emission of radiation were discrete processes.

It was still generally believed that radiation propagated as a wave.

This all changed in 1905 when Einstein published his revolutionary explanation of the photoelectric effect.

2.2.1 The Photoelectric Effect - The Problem

When Hertz first succeeded in producing electromagnetic waves from oscillating electrical circuits, he noted that "it is essential that the pole surfaces of the spark gap [from which the waves were emitted] should be frequently re-polished".

He also observed that sparks were more readily induced between the poles in the secondary (detector) circuit when they were illuminated with ultra-violet radiation.

He called this phenomenon, for which he had no explanation, the *Photoelectric Effect*.

A detailed study of the photoelectric effect was subsequently made by Hallwachs.

He studied the effect of the following three factors on the response of a charged gold-leaf electroscope to visible light and ultra-violet radiation (Fig 2.14).

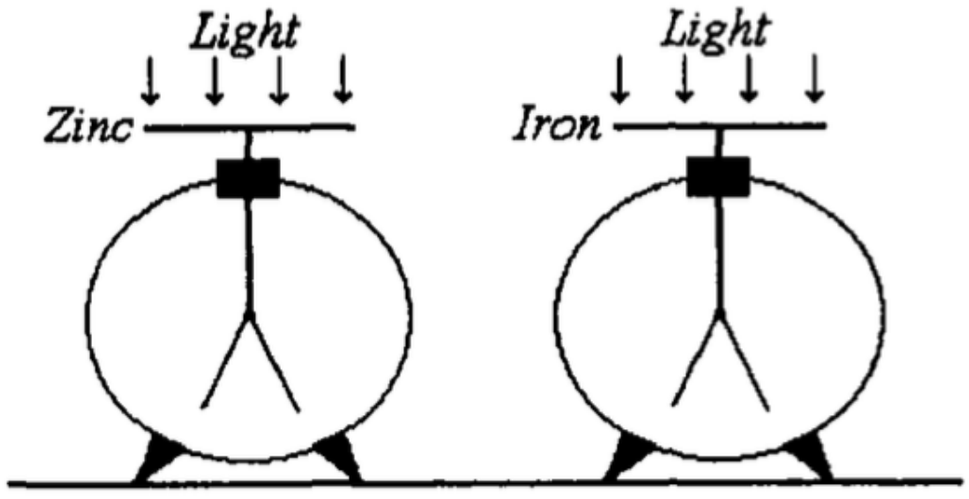


Fig 2.14 Hallwach's experiment. Two gold-leaf electroscopes, one with an iron plate the other with a zinc plate are charged, first positively and then negatively. In each *case*, light of different colors is beamed on to the plates.

1. **The sign of the charge carried by the electroscope, positive or negative.**
2. **The metal from which the electroscope plate was made, iron or zinc.**
3. **The intensity of the light and ultra-violet radiation.**

Hallwachs made the following observations:

* **Illuminating the plate of a positively charged electroscope with visible light or ultra-violet radiation produced no response whatsoever, whether the plate was made of iron or zinc. Even if a very intense beam of radiation was used, no charge was lost by the positively charged electroscope.**
* A negatively charged iron plate electroscope showed no response when its plate was illuminated by an intense beam of either visible light or of ultra-violet radiation.
* Visible light also had no effect on a negatively charged zinc plate electroscope. However, the moment the plate of the negatively charged zinc plate electroscope was illuminated with even a weak beam of ultra-violet radiation, the electroscope's leaves closed, indicating that it had lost its negative charge.

The discharge of the negatively charged zinc plate electroscope is readily explained by supposing that the ultra-violet radiation supplies the energy needed by the electrons constituting its negative charge to break the bond holding them to the zinc plate and to escape from it.

However, this still leaves a number of unanswered questions:

1. **Why are no electrons released from the iron plate electroscope by the ultra-violet radiation? Why the difference in the behaviour of the two metals? All metals emit electrons when heated—the *thermionic effect*. Why does irradiating metals with light energy affect them differently from warming them with heat energy?**

2. Whereas ultra-violet radiation of any intensity, no matter how weak, instantaneously releases electrons from the negatively charged zinc plate, electrons are not released from either plate by visible light, no matter how intense the beam. Why is this? The intensity of an electromagnetic wave, namely, the rate at which it delivers energy, is a function of its amplitude, not of its frequency. Thus, a radiation's colour (its frequency) should not be a factor in its ability to deliver the energy needed for the release of the electrons from the metal plates.
   Furthermore, the greater the intensity of the ultra-violet radiation, the greater should be the kinetic energy of the liberated electrons—the *photoelectrons*. However, other experiments showed that although the intensity of the ultra-violet radiation does affect the number of electrons emitted from the zinc plate—the more intense the beam the more electrons are released—it has no effect on their individual energy.
3. The energy carried by a wave is distributed equally along its front. Thus, the energy of the incident electromagnetic wave should be delivered at the same rate to every electron in the surface of the metal plate and a quite considerable length of time should pass before they have accumulated sufficient energy to be released, whereupon they should all be emitted in a single explosive burst. Yet the zinc plate electroscope loses its negative charge the very moment its plate is illuminated with ultra-violet radiation, even at low beam intensities. How does this happen?

Classical electromagnetic theory had no answer to these questions.

Evidently, the photoelectric effect requires a totally new theory of the transport of energy by radiation.

2.2.2 Einstein's Equation

The photoelectric effect was first elucidated by Einstein in 1905.

In doing so, he adopted and extended Planck' quantum hypothesis, postulating that "the incident light is composed of quanta of energy"

Einstein proposed that radiation is not only absorbed and emitted in quanta but that it also propagates as such.

Thus, a light beam can be considered to be a stream of particles, *light quanta* or *photons,* each of energy

$$E_{Photon} = hf \tag{2.17}$$

where f is the frequency of the light.

The photons move through space in straight lines and at the speed of light.

Light intensity falls as we move away from a light source because the photons it emits are spread out over a larger surface area.

However, the energy of each individual photon remains unchanged.

Having ascribed a particle nature to light, Einstein went on to treat the photoelectric effect as a simple mechanical problem of energy conservation.

> ***The quanta penetrate the surface of the material and their respective energies are at least in part changed into the kinetic energy of the electrons ... The simplest process conceivable is that a quantum of light gives up all its energy to a single electron ...Upon reaching the surface, an electron originally inside the body will have lost a part of its kinetic energy. Furthermore, one may assume that each electron in leaving the body does an amount of work which is characteristic of the material. Those electrons that are ejected normal to and from the immediate surface will have the greatest velocities.***[20]

A photon striking a metal surface surrenders all of its energy to a single electron in the metal.

The electron first uses this energy to break the bond holding it to the metal; any surplus energy appears as the kinetic energy, $K_{Electron}$, of the liberated electron.

Einstein summarized these ideas in the following simple relationship called *Einstein*'s *Equation:*

$$\begin{aligned} K_{Electron} &= E_{Photon} - W \\ &= hf - W \end{aligned} \tag{2.18}$$

W is the work done in releasing the electron from the metal plate.

The minimum amount of work, W_{min}, required to release an electron from a metal plate is called the metal's work *function;* it is a characteristic property of the metal.

Einstein's simple hypothesis explains the results of the experiments with the charged electroscopes as follows:

A. **Ultra-violet radiation does not liberate electrons from an iron plate because the metal's work function is greater than the energy of the photons, i.e., $W_{min} > hf_{UV}$. In order to release electrons from iron, radiation of higher frequency must be used; radiation whose photons have a greater individual energy.**

B. **Visible light comprises a range of frequencies - those of all the colours of the rainbow. Thus, it contains photons in a range of energies, from about 2 to 3** electron-volts. However, the work function of zinc (and certainly that of iron) is greater than the energy of even the most energetic of these photons. Hence, visible light cannot release electrons from a zinc plate.

C. Electrons are released the moment the ultra-violet radiation strikes the zinc plate because the energy concentrated in the photons is transferred to them virtually instantaneously. There is no accumulation of energy.

D. The photons in an intense beam of radiation possess no more individual energy than those in a weak beam of the same radiation; the more intense beam just contains more photons. Thus, an intense beam can liberate more photoelectrons from the metal plate. However, not all the photoelectrons are released with the same kinetic energy; those extracted from deep inside the metal plate have to expend some of their energy in rising to the surface. The photoelectrons with the greatest kinetic energy, K_{max}, are those released from the plate's surface, since these have to expend no more than the metal's work function, W_{min}, in breaking loose.

$$K_{max} = hf - W_{min} \tag{2.19}$$

The color (frequency) dependence of the response of certain materials to incident light had in fact been known since the invention of photography, some sixty years before the discovery of the photoelectric effect.

It was well known that most photographic plates were not sensitive to red light; red and black objects looked the same in photographs.

Furthermore, the plates could be safely exposed and developed in a room illuminated with red light.

On the other hand, even very weak blue light ruined the plates.

The imprinting of an photographic image on a photographic plate or film involves a chemical reaction, the energy for which is supplied by light; such reactions are called *photochemical reactions*.

The reaction involved in photography is essentially the reduction of silver ions to metallic silver:

$$Ag^+ + e^- \xrightarrow{\text{photon}} Ag$$

Red light, which has a relatively low frequency, did not blacken the plates used by the first photographers because its photons have insufficient energy to promote the reduction of the silver ions.

By the addition of suitable dye-forming chemicals to the emulsion containing the silver ions, modem photographic films are made sensitive to all the colors of the rainbow and even to infra-red radiation.

Einstein's proposal that light propagates as photons stood in total opposition to electromagnetic wave theory.

The fundamental concept implicit in Einstein's revolutionary *quantum theory of light* was that photons are real particles, a notion totally alien to the classical wave theory of light.

There was no way of arriving at the concept of photons from Maxwell's equations.

Moreover, it was subsequently shown that the function derived by Planck for the energy distribution between the wavelengths in black-body radiation function could also be derived by regarding cavity radiation as a sort of photon gas between whose particles the radiation's energy was suitably distributed.

The real nature of light was once again an open question.

2.2.3 Planck'sConstant

Planck's constant, *h,* is one of the four fundamental constants of physics.

The other three are the velocity of light ***i***n *vacuo, c,* the universal gravitational constant, G, and the elementary charge, *e*. Planck's constant is the underlying constant of quantum physics.

Because of its very small value, $h = 6.63 \times 10^{-34}$ J s, quantum effects are not usually detected on the macroscopic scale.

However, if its value was zero, quantum effects would not exist at all.

The three fundamental constants, h, c, and e, can be combined with the permittivity of empty space, ε_0, to give a dimensionless number, $\alpha = e^2/2\varepsilon_0 hc \approx 1/137$ called the *Fine Structure Constant*.

This number is central to the theory of Quantum Electrodynamics - QED, that unites quantum theory and the theory of relativity in explaining the interaction of photons and electrons.

In 1916, Millikan designed and carried out an accurate experiment that both confirmed the validity of Einstein's equation and fixed the precise value of Planck's constant.

The key component of the experiment was a glass vacuum tube called a photoelectric tube or photoelectric diode, into which two electrodes were fixed: one, a metal plate that emitted electrons when illuminated by visible light; the other smaller, intended for the collection of the photoelectrons.

The electrodes were connected to a potentiometer, such that the illuminated metal plate was given a positive potential and the collector a negative potential (Fig 2.17).

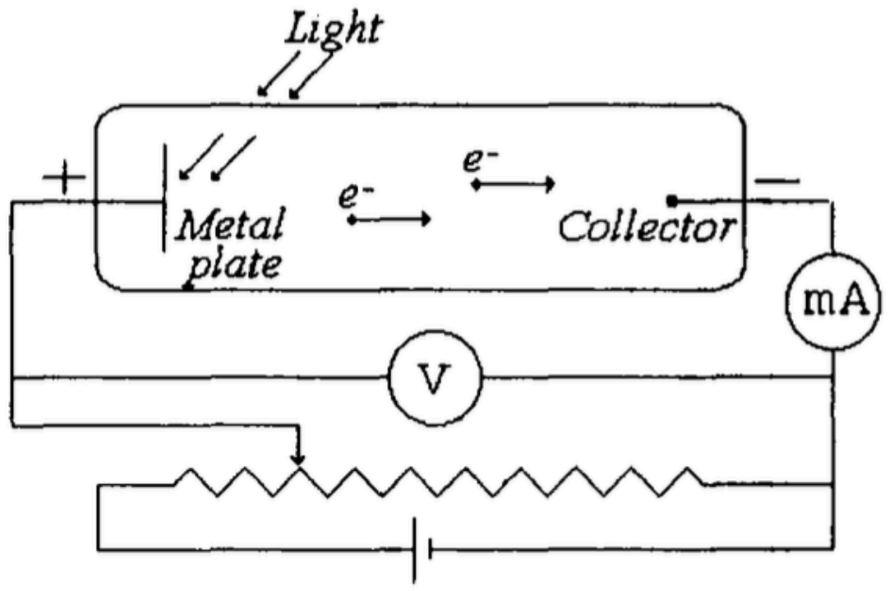


Fig 2.17 Millikan's apparatus used to investigate the photoelectric effect. Light falls on the metal plate and the photoelectrons emitted from it are collected by the electrode. The potential difference, V, between the electrodes can be adjusted by the potentiometer.

The physical quantity that Millikan measured in the experiment was the kinetic energy of the most energetic photoelectrons released from a metal by incident light of a given frequency,

In order to reach the collector electrode, an electron emitted from the illuminated metal plate has to move against the potential difference, V, between the two electrodes.

It does so at the expense of its kinetic energy.

As long as photoelectrons manage to overcome the potential difference and reach the collecting electrode, an electric current, whose magnitude is shown by the ammeter reading, flows through the circuit.

In order to measure the kinetic energy of the most energetic photoelectrons released by the incident light, Millikan determined the minimum electrical potential - the stopping potential - needed to reduce the current to zero.

The metal plate in the photoelectric tube was illuminated with monochromatic light of known frequency, f .

The potential between the electrodes was adjusted by means of the potentiometer until no current passed through the ammeter.

The stopping potential, V_0, was then recorded.

The experiment was repeated with light of different frequencies and with plates made of different metals.

The change in an electron's kinetic energy, ΔK, as it moves through a potential difference, V, is given by:

$$\Delta K = eV \tag{2.20}$$

Thus, the stopping potential V_0 is related to the maximum kinetic energy, K_{max} possessed by the photoelectrons by the equation

$$K_{max} = eV_0 \tag{2.21}$$

Substituting this expression in Einstein's equation -equation (2.17)- gives:

$$eV_0 = hf - W_{min} \tag{2.22}$$

or, alternatively

$$V_0 = \frac{hf}{e} - \frac{W_{min}}{e} \tag{2.23}$$

This function, which was derived directly from Einstein's equation, predicts a linear relationship between the stopping potential, V_0, and the frequency, f of the light.

This was confirmed by Millikan's experiment.

Typical experimental results for three different metals are shown in Fig 2.16.

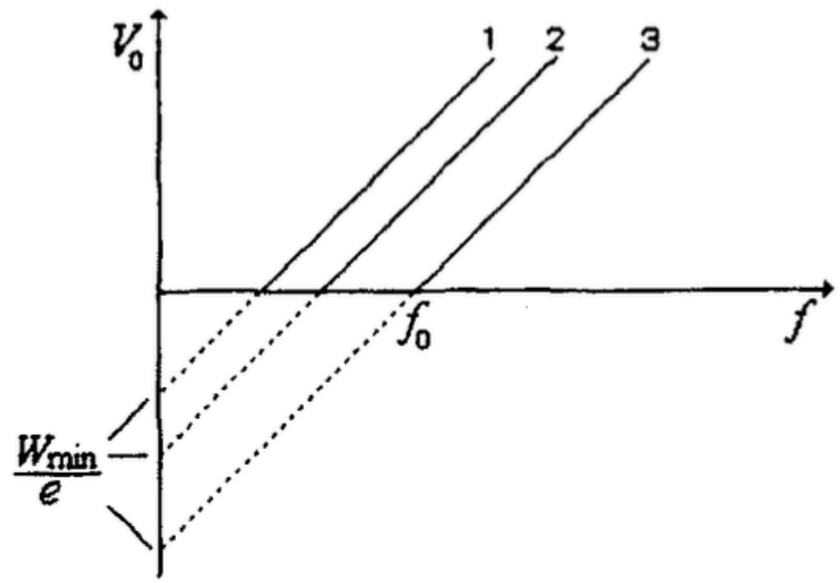


Fig 2.18 The stopping potential, V_0, as a function of the frequency, f , of the light with which they are irradiated for three different metals. The metals' work functions, W_{min}, can be calculated from the points where the extrapolated straight lines cut the vertical axis.

Planck's constant can be calculated from the slope, h/e of the lines.

The point at which each line cuts the horizontal axis gives the minimum frequency - the threshold frequency, f_0, which releases electrons from the metal.

From the line's intersect with the vertical axis, the value of work function, W_{min}, of each metal can be obtained.

2.2.4 X-rays

In the photoelectric effect, photons are annihilated when they strike the metal plate, each one surrendering its energy to an electron.

Is the reverse process possible?

Can the kinetic energy of electrons be used to produce photons?

This reverse process had been discovered, but incorrectly interpreted, a number of years before Planck put forward his quantum hypothesis.

In 1895, Roentgen had found that an invisible radiation of great penetrating power was produced when the "electrical discharge from a fairly large induction coil" was passed through a cathode ray tube.

The most striking feature of this phenomenon is the fact that an active agent [the radiation] here passes through a black cardboard envelope, which is opaque to the visible and ultra-violet rays of the sun ... an agent which has the power of producing active fluorescence.
We soon discover that all bodies are transparent to this agent, though in very different degrees ... If the hand be held between the discharge tube and the screen, the darker shadow of the bones is seen within the slightly dark shadow image of the hand itself.[23]

Roentgen called this radiation *X-rays*.

He found that X-rays propagate in straight lines from which they are deflected neither by electric nor magnetic fields.

They ionize the matter through which they pass, cause fluorescent materials to glow and darken photographic plates.

X-rays are now known to be high frequency electromagnetic radiation, $l0^{17}$Hz to 10^{20}Hz.

They show all the characteristic wave properties of electromagnetic radiation such as diffraction and polarization.

However, their refractive index is almost exactIy equal to unity in all materials and so they cannot be focused by lenses.

Fig 2.17 shows the basic design of a typical device for producing X-rays.

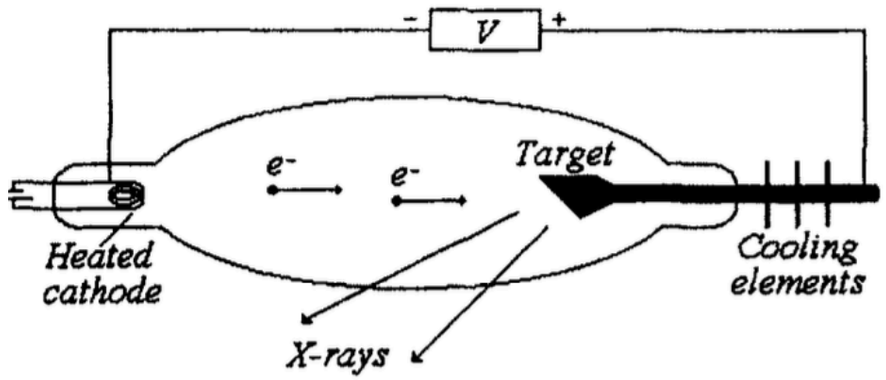


Fig 2.17 An X-ray tube. Electrons emitted from hot cathode are accelerated towards the target by the potential difference *V*. X-rays are produced from the target.

Electrons, accelerated by potential differences of tens of thousands of volts, are aimed at a metal target in a vacuum tube.

Upon striking the target, the electrons are abruptly stopped and X-rays and heat are generated.

Just 1% of the kinetic energy of the electrons is converted into X-rays; the rest appears as heat.

Consequently, the target must be cooled when the tube is operated.

Typical intensity spectra of the X-ray radiation produced at various potential differences by such devices are shown in Fig 2.18.

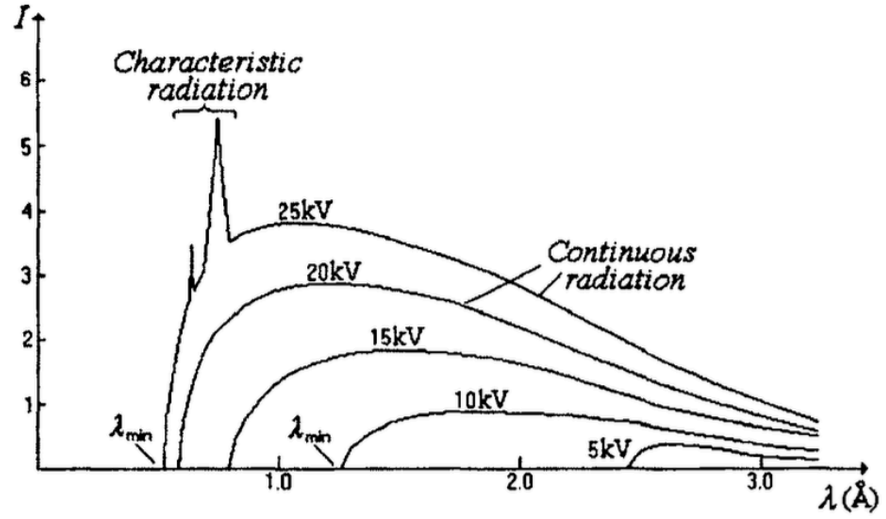


Fig 2.18 The intensity, I, of the radiation produced from a typical X-ray tube (molybdenum target) as a function of its wavelength, λ, for different values of the potential difference at which the tube is operated. Note the characteristic radiation which appears as sharp peaks of intensity on the spectrum of the continuous radiation at high potential differences.

All the spectra exhibit a sharp cut-off at a specific wavelength below which radiation is not produced.

This minimum wavelength, λ_{min} is inversely proportional to the potential difference,V, at which the device is operated.

At low and medium potential differences, 5kV to **2**0kV, the spectrum above the cut-off is continuous.

Changing the material from which the target in the tube is made has no effect on the spectrum of this *continuous radiation*.

At high potential differences, >20kV, sharp intensity peaks appear in the spectrum at wavelengths that are a characteristic of the material from which the anode target is made.

Initially, the origin of the continuous radiation was accounted for in classical terms.

According to classical electromagnetic theory, all accelerated or decelerated charges radiate electromagnetic waves.

That produced by decelerated charges was called bremstrahlung.

It was suggested that, on striking the target in the X-ray tube, the electrons are decelerated and that this deceleration was the source of the continuous bremstrahlung radiation.

However, no explanation could be found for why radiation is only produced down to a certain minimum wavelength nor why this minimum wavelength depends directly on the potential at which the X-ray tube is operated.

In contrast, these features are readily explained by the quantum theory of radiation.

Although most of the electrons striking the target are decelerated in a series of multiple collisions with the target's metal atoms, a few of them may lose all or almost all of their energy in a single collision.

The electrons that undergo multiple collisions generate the heat produced when the tube is operated; those stopped by a single collision generate the X-rays.

Thus, the production of the X-rays is a sort of reverse photoelectric effect.

In the simplest case, an electron loses all its energy in a single event, generating a single X-ray photon with a corresponding energy.

The kinetic energy of an electron accelerated from rest by a potential difference V is $K_{Electron} = eV$.

The most energetic photon that can be produced from such an electron will possess all this energy:

$$\left(E_{Photon}\right)_{\max} = eV = hf_{\max} \tag{2.24}$$

The wavelength of the radiation associated with these photons will be

$$\lambda_{\min} = \frac{c}{f_{\max}} = \frac{ch}{eV} \tag{2.25}$$

i.e., the minimum wavelength produced by the electrons is inversely proportional to the accelerating potential, V, as is found experimentally.

2.2.5 X-rays and Crystallography

The nature of X-rays was the subject of much controversy during the years immediately following their discovery.

Were they high energy cathode rays or a wave phenomenon?

And if they were waves, what sort were they?

At one point, Roentgen even suggested they might be *longitudinal* aether waves.

The wave nature of X-rays was ultimately established in 1912, when von Laue suggested that crystals could be used as diffraction gratings for X-rays.

Diffraction occurs whenever the progress of a wave-front is restricted by an object; the extent and nature of the diffraction depending on the ratio between the wavelength and the size of the object.

When the spacing between a grating's elements is of the order of magnitude of the wavelength of the radiation incident upon it, the grating diffracts the radiation in certain preferred directions.

Von Laue anticipated that the spacing between the atoms in crystals would be of the order of magnitude expected for X-rays.

The idea was tried and diffraction patterns were indeed obtained.

The result was viewed as confirmation of the wave nature of X-rays.

In 1913, William Bragg and his son Lawrence proposed a simple method of measuring the wavelengths of X-rays by means of their ‘reflection’ from crystal planes.

Crystals comprise atoms (or ions) that are arranged in a spatial lattice, such that they define ‘families’ of parallel planes (Fig 2.19).

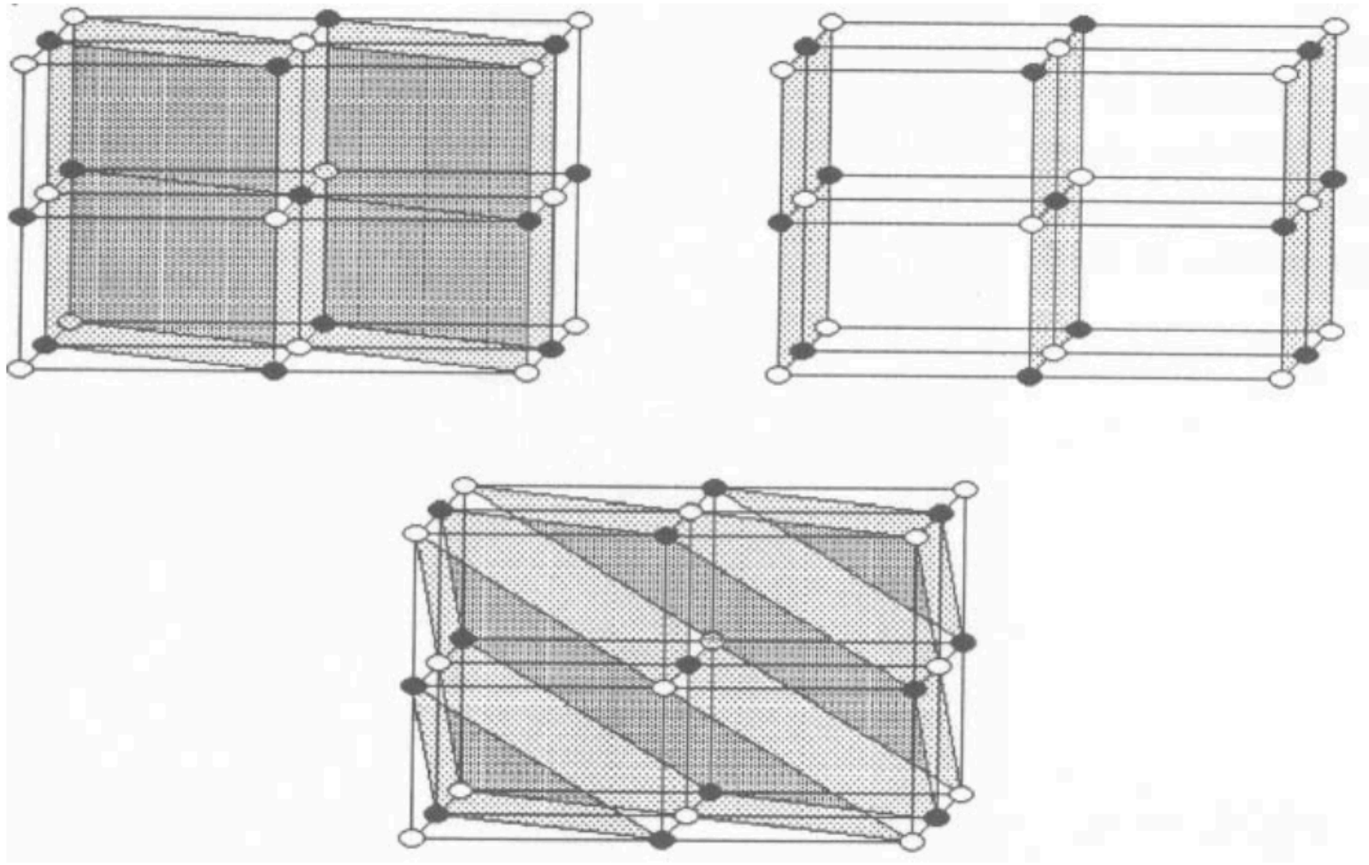


Fig 2.19 ‘Families’ of parallel planes in the same crystal. In each family the distance between the planes is different.

The Braggs proposed that beams of X-rays would be reflected from these planes as though they were plane mirrors (Fig 2.20).

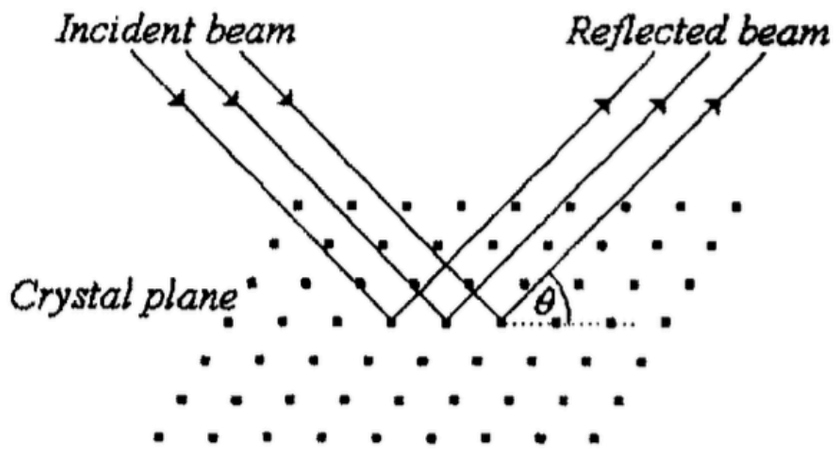


Fig 2.20 The 'reflection' of a beam of electromagnetic radiation from a crystal plane. The glancing angle θ is that between the crystal plane and both the incident and the reflected beam.

The reflection of X-rays from a crystal plane can be understood in terms of Huygens' principle, in the same way as it explains the reflection of a beam of light from a plane mirror.

According to this principle, when a wave-front strikes a reflecting surface, each point on the surface acts as a source of secondary light waves.

The secondary waves interfere constructively only where the angle of reflection of the beam equals its angle of incidence such that the new wavefront is the common tangent to these secondary waves (Fig 2.21).

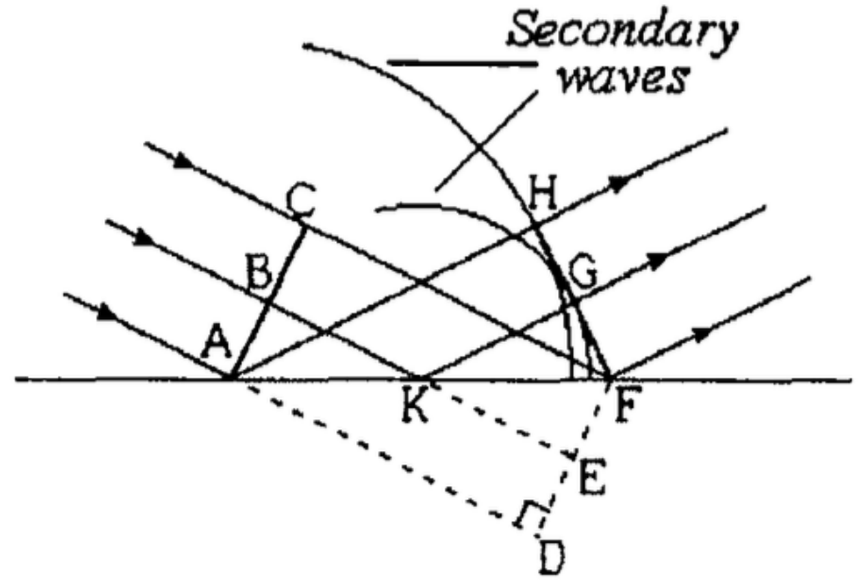


Fig 2.21 The plane wave-front ABC represents a parallel beam of light which strikes a plane mirror obliquely. In the absence of the mirror, the wave-front would have travelled to DEF. Instead, it is reflected and occupies the position FGH. According to Huygens principle, the new wave-front is tangential to the secondary waves whose centers are at the points A and K. The right-angled triangles KGF, KBA and KEF are congruent; so too are the triangles AHF, ACF and ADF. Hence the incident and reflected wave-fronts are equally inclined to the mirror.

Similarly, when a beam of X-rays is aimed at a crystal, its atoms act as sources of secondary waves and the beam is reflected at an angle equal to its angle of incidence.

The X-rays reflected from a family of parallel planes all emerge from the crystal at the same glancing angle θ relative to the planes.

However, the distances the rays have travelled inside the crystal may differ, depending on the depth of the plane from which they were reflected.

Whether these rays interfere constructively or destructively when they are recombined in the reflected beam depends on their path difference.

Consider rays reflected from two adjacent planes spaced at a distance d (Fig 2.22).

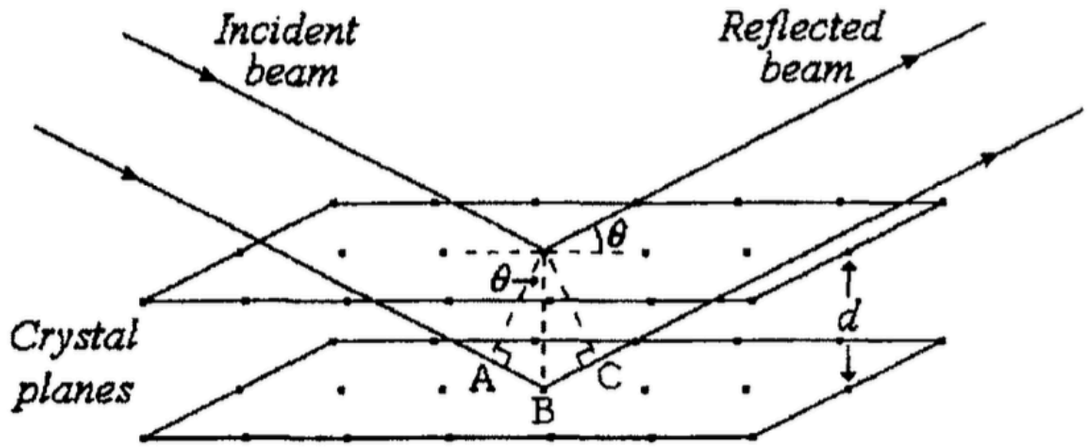


Fig 2.22 The reflection of X-rays from two adjacent parallel crystal planes. The optical path of the ray reflected from the lower plane is longer than that reflected from the upper plane.

The path difference between the rays is

$$\mathrm{AB} + \mathrm{BC} = 2d\sin\theta \tag{2.26}$$

Destructive interference occurs when the path difference equals an odd number of half the radiation's wavelength:

$$(2n+1)\lambda/2 = 2d\sin\theta_n \qquad (n = 1, 2, 3, \ldots) \tag{2.27}$$

at these angles, the reflected beam will then be of minimum intensity.

The reflected beam of radiation will be of minimum intensity at glancing angles given by

$$\theta_n = \sin^{-1}\left((2n+1)\lambda/4d\right)$$

Constructive interference occurs when the path difference equals an integer number of the radiation's wavelength:

$$n\lambda = 2d\sin\theta_n \quad (n = 1, 2, 3, \ldots) \tag{2.28}$$

the reflected beam will then be of maximum intensity; n is termed the order of the reflected beam.

The reflected beam of radiation will be of maximum intensity at glancing angles given by

$$\theta_n = \sin^{-1}\left(n\lambda/2d\right)$$

Thus, when a crystal is rotated in a beam of X-rays, the intensity of the beam reflected from it rises and falls as the angle at which the beam strikes the crystal changes.

Equation (2.28) is known as *Bragg's Equation*.

If the wavelength of the radiation is known, X-ray crystallography can be employed to determine the structure of crystals and to measure interatomic distances in molecules.

However, its use is limited to regularly repeating structures, as only these can give rise to constructive interference in the reflected or diffracted beam.

The structure of many proteins has been elucidated by this method; the helical structure of DNA was first indicated by the X-ray diffraction pattern to which it gave rise.

Part 2.3 Photons

Einstein presented his quantum theory of light and its elucidation of the photoelectric effect in a scientific paper that appeared in the spring of 1905.

He followed it, just a few weeks later, with a ground-breaking paper on the mathematics of Brownian motion.

His special theory of relativity was published in the summer of that *annus mirabilis* and before the year's end he had also completed his doctoral thesis.

In the space of just a few months, Einstein had set much of the agenda of physics for the next 50 years.

2.3.1 Photon Mass

Now we will examine some of the properties of the light quanta -the *photons*- whose existence Einstein first conceived.

$$E_{photon} = hf \tag{2.29}$$

The quantum theory of light postulates that electromagnetic radiation propagates as particles -photons- that travel at the speed of light.

According to this theory, the energy possessed by a photon depends only on the frequency, f; of the radiation with which it is associated:

However, as we saw in (Part 1.4), according to the special theory of relativity, the total energy, E, of any particle is given by

$$E^2 = m_0^2 c^4 + c^2 p^2 \tag{1.67}$$

where m_0 is the rest-mass of the particle, p is its linear momentum and c is the velocity of light in *vucuo*.

How can these two definitions of the particle's energy be reconciled in the case of the photon?

According to the principle of the constancy of the speed of light, no inertial observer can ever catch up with a photon in free space and move together with it such that it appears to be at rest; photons are never at rest.

Consequently, the concept of the 'rest-mass of a photon' has no meaning; the rest-mass of a photon can be taken to be zero.

Substituting the photon's zero rest-mass, $m_0 = 0$, in equation (1.67) gives

$$E_{photon}^2 = c^2 p_{photon}^2 \tag{2.30}$$

from which we obtain:

$$p_{photon} = \frac{E_{photon}}{c} \tag{2.31}$$

Thus, even though a photon has no rest-mass, a linear momentum, p_{photon}, can be attributed to it.

Substituting $E_{photon} = hf$ and $c = f\lambda$, in equation 2.31 gives

$$p_{photon} = \frac{hf}{c} = \frac{h}{\lambda} \tag{2.32}$$

where λ is the radiation's wavelength in *vacuo*.

This simple relationship between the photon's linear momentum, p_{photon}, and the wavelength, λ, or frequency, f of the radiation with which it is associated, is the fundamental link between the quantum theory of light and the wave theory of light.

Although photons have no rest-mass, it is often useful to attribute a moving mass to them.

Equating the relativistic expression for the total energy of a particle, $E = mc^2$,with that for the energy of a photon, $E = hf$, gives:

$$m_{photon} = {hf}/{c^2} \tag{2.33}$$

In an article published in 1911, Einstein predicted that as a consequence of photon mass, a ray of light passing through the sun's gravitational field close to its surface would be deflected through an angle of 0.00025° (0.9 seconds of arc).

Similar predictions had been previously made using Newtonian mechanics, by calculating the deflection of an ordinary particle that passed close to the sun at the speed of light.

Neither of these calculations took account of the curvature of space -its non-Euclidean nature- predicted by the theory of general relativity.

In 1915, Einstein repeated his calculations to allow for space curvature and predicted a deviation of 0.0005° (1.8 seconds of arc).

This prediction was confirmed during the total solar eclipse in 1919 when it was found that a star whose line of sight passed close to the sun's surface appeared to be displaced from its true astronomical position (Fig. 2.26).

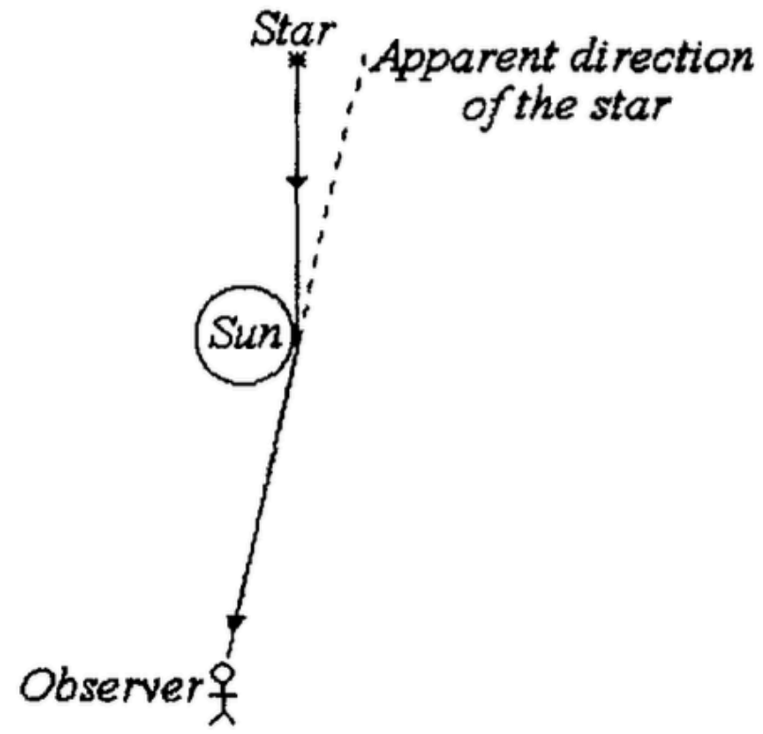


Fig.2.24 The deflection of a ray of starlight by the sun's gravitational field. During a total solar eclipse, when the sun's glare does not blind the observers, the star can be viewed and appears to be displaced from its true position. The deviation is attributed to the bending of the light ray as it glances the sun's surface.

The observed deviation was 0.00045° ± 25%. This result was received as evidence of the gravitational mass of the photon and as a confirmation of the theory of general relativity.

Photons rising from the earth's surface lose energy as a result of the work that must be done against the earth's gravitational pull.

This loss does not appear as a reduction in their speed-photons always travel at the speed of light-but as a reduction in the frequency of the radiation.

The phenomenon is called the gravitational red-shift because, in the case of visible light, the frequency of the rising photons is shifted in the direction of the red end of the spectrum.

This reduction in frequency was observed in experiments conducted in 1960, employing the Mossbauer effect.

Like any stream of particles, a beam of photons exerts a force on any surface it, discovered even before the strikes.

The phenomenon, called *radiation pressure* was a development of the quantum theory.

It was found that a small light paddle-wheel could be made to rotate by aiming a beam of light at it.

The phenomenon was explained at the time in terms of classical electromagnetism.

It was reasoned that the force exerted by the radiation resulted from the interaction of the electrical and magnetic fields in the incident electromagnetic waves with the electrons in the surface of the paddle-wheel.

It was shown that the linear momentum transferred by the incident radiation to the Surface, $p_{radiation}$, is related to the energy it delivers, $E_{radiation}$, by

$$p_{radiation} = \frac{E_{radiation}}{c} \tag{2.34}$$

The similarity between this relationship derived from classical electromagnetic considerations and that derived from relativistic considerations, equation (2.31), is evident and is suggestive of a consistency underlying electromagnetic theory, relativity and the quantum theory of radiation.

2.3.2 The Compton Effect

Confirmation of the quantum theory of light and, in particular, of its prediction that a linear momentum can be attributed to photons, was obtained in a landmark experiment carried out by Compton in 1923.

The experiment set out to investigate the scattering of high energy photons -X-ray photons- by free electrons.

According to classical electromagnetic theory, radiation incident on free electrons should be scattered in all directions, without undergoing any change in its frequency.

The incident waves set up vibrations in the electrons at frequencies equal to their own, and these, in turn, generate the secondary waves of the scattered radiation, also at the same frequencies.

Overall, the free electrons should not take up any energy from the incident radiation.

Relativistic considerations appear to support this conclusion.

Conservation of linear momentum requires that a free stationary electron that absorbs an amount of energy, $E_{radiation} = pc$, from an incident beam of electromagnetic radiation, must also acquire a corresponding amount of linear momentum, $p = E_{radiation}/c$.

The energy absorbed by the electron appears as kinetic energy, K, such that

$$K = E_{radiation} = pc \tag{2.35}$$

However, according to special relativity, the kinetic energy of a moving particle is given by

$$\begin{aligned} K &= mc^2 - m_0c^2 \\ &= \sqrt{m_0^2c^4 + p^2c^2} - m_0c^2 \end{aligned} \tag{2.36}$$

Clearly, this relativistic expression is incompatible with the conservation requirement given by equation (2.35) that K = pc.

Thus, it would appear, that for free electrons to absorb the energy of incident electromagnetic radiation would require a violation of the principles of energy and linear momentum conservation.

We may note, in passing, that this problem does not arise in the photoelectric effect, where the electrons taking up energy from the radiation are bound to the atoms in the metal plate.

In this case, the atoms participate together with the electrons in absorbing the energy and momentum of the incident radiation.

The heavy particles absorb the marginal amounts of momentum or energy not absorbed by the electrons, thereby ensuring momentum and energy conservation.

However, notwithstanding all these considerations, free electrons do in fact absorb energy from high frequency electromagnetic radiation.

When a beam of X-rays of frequency, f , passes through a medium containing free electrons, X-rays of lower frequencies, $f' < f$, are detected in the emergent beam.

In terms of the quantum theory of light, the lower frequencies indicate the presence of lower energy photons in the emergent beam.

This implies that photons in the incident beam must have surrendered energy during the passage of the radiation through the scattering medium.

Apparently, an interaction does occur between the incident X-ray photons and the free electrons in which energy is transferred from the former to the latter.

Compton suggested that this interaction could be elucidated by means of an amalgam of classical mechanics, the quantum hypothesis and special relativity.

According to Compton, photons of the incident X-rays collide elastically with the free electrons.

After colliding, the photons and electrons move off in different directions. (Fig. 2.25).

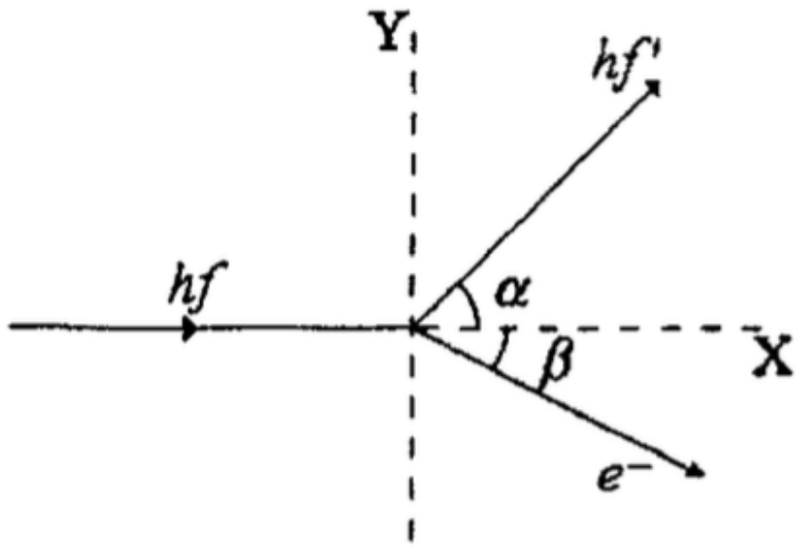


Fig. 2.25 An elastic collision between a photon of frequency f and a stationary free electron. The electron recoils at an angle β relative to the direction of the incident photon.

The energy of the photon after the collision, hf', is less than that of the incident photon, hf the difference, h($f - f'$) appears as the kinetic energy, K, of the electron:

$$K = h(f - f') \tag{2.37}$$

The total (relativistic) energy, E, of the electron after the collision is equal to the sum of its rest energy, m_0c^2, and its kinetic energy, K:

$$\begin{aligned} E &= m_0c^2 + K \\ &= m_0c^2 + h(f - f') \end{aligned} \tag{2.38}$$

According to the quantum theory of radiation, the relationship between the linear momentum of a photon and the frequency of its associated radiation is given by $hf = cp_{photon}$.

Substitution in equation (2.38) gives

$$E = m_0c^2 + c\left(p_{photon} - p'_{photon}\right) \tag{2.39}$$

where p_{photon} is the magnitude of the photon's linear momentum before the collision and p'_{photon} is its magnitude afterwards.

A second expression for the electron's total energy after the collision can be obtained from the equation for the total energy of a particle given by special relativity:

$$E^2 = m_0^2c^4 + c^2p_{electron}^2 \tag{2.40}$$

where m_0 is the rest-mass of the electron and $p_{electron}$ is the magnitude of its linear momentum after the collision.

Combining the two expressions for the electron's energy - equations (2.39) and (2.40)- gives:

$$m_0^2c^4 + c^2p_{electron}^2 = \left[m_0c^2 + c\left(p_{photon} - p'_{photon}\right)\right]^2 \tag{2.41}$$

After the collision, the original linear momentum, $\vec{p}_{photon}$, of the incident photon will be divided between that, $\vec{p}_{electron}$, of the recoiling electron and that, $\vec{p}'_{photon}$, of the scattered photon:

$$\vec{\mathbf{p}}_{photon} = \vec{\mathbf{p}}'_{photon} + \vec{\mathbf{p}}_{electron} \tag{2.42}$$

Separating this vector equation into components gives:

$$p_{photon} = p'_{photon}\cos\alpha + p_{electron}\cos\beta \qquad \text{(X axis)} \tag{2.43}$$

$$0 = p'_{photon}\sin\alpha + p_{electron}\sin\beta \qquad \text{(Y axis)} \tag{2.44}$$

where α and β are, respectively, the scattering angles of the photon and electron.

By applying the general trigonometric relationship $\sin^2\theta + \cos^2\theta = 1$ and eliminating β from these two equations we obtain:

$$p^2_{electron} = p^2 - 2p_{photon}p'_{photon}\cos\alpha + {p'_{photon}}^2$$

Substituting $p^2_{electron}$ from equation (2.45) in equation (2.41) and rearranging gives (2.45)

$$\frac{1}{p'_{photon}} - \frac{1}{p_{photon}} = \frac{(1-\cos\alpha)}{m_0 c} \tag{2.46}$$

This relationship can be re-written in terms of the difference, $\lambda' - \lambda$, between the respective wavelengths of the emergent and incident radiation by the substitutions

$$\frac{1}{p'_{photon}} = \frac{\lambda'}{h}$$

and

$$\frac{1}{p_{photon}} = \frac{\lambda}{h}$$

from equation (2.32):

$$\lambda' - \lambda = \frac{h}{m_0 c}(1 - \cos\alpha) \qquad (2.47)$$

Thus, Compton predicted that the reduction in the radiation's wavelength should depend only on the scattering angle; α, all the other factors, h, m_0 and c, appearing in equation (2.47) being constants.

In the experiment he conducted to test this prediction, Compton directed a beam of X-rays onto a graphite target.

Since the binding energy of the outermost electrons in the carbon atoms that comprised the target is negligible compared to that of the high energy X-ray photons, these electrons could be considered to be essentially free.

Using the Bragg technique, Compton compared the wavelength of the incident X-rays with that of the scattered X-rays at specific angles, 45°, 90° and 135°.

In each case, he found that the difference between the wavelength of the scattered X-rays and that of the incident X-rays, $\lambda' - \lambda$, corresponded to that predicted by equation (2.47).

The phenomenon is called the *Compton effect*.

Shortly afterwards, Wilson and Bothe succeeded in tracking the recoiled electrons by the trails they left in a cloud chamber.

Since the increase in the wavelength of the scattered radiation, $\Delta\lambda = \lambda' - \lambda$, is independent of the wavelength of the incident radiation, the fractional change, $\Delta\lambda/\lambda$, becomes negligible at long wavelengths.

Thus, the Compton effect is only exhibited by short-wave radiation: X-rays exhibit a maximum fractional change of about 10^{-2}; γ-rays of the order of one.

Furthermore, because the mass of the scattering particle, m_0, appears in the denominator of equation (2.47), increases in the wavelength of even high-energy radiation axe only detected when the scattering entities are as small as electrons.

The change in wavelength caused by a scattering particle as large as an atom is totally negligible.

Thus, no detectable change occurs in the wavelength of the X-rays scattered by the atoms and ions comprising the crystal lattice in the Bragg technique.

In terms of quantum theory, the mechanism of the Compton effect can be regarded as a sequence of two electron-photon interactions:

* **In the first, the free electron absorbs energy $E = hf$ and linear momentum of magnitude $p = h/\lambda$ from the incident photon.**
* **In the second, a photon with energy $E' = hf'$ and linear momentum of magnitude $p' = h/\lambda'$ is emitted from the electron.**

Overall, the electron acquires amounts of energy and linear momentum of magnitudes ΔE = E- E' and Δp = p - p', respectively, such that they satisfy the relativistic requirement $\Delta E = \sqrt{m_0^2 c^4 + (\Delta p)^2 c^2} - m_0 c^2$.

The Compton effect is not the only phenomenon amenable to interpretation as a sequence of particle and photon interactions.

All the interactions between charged particles described by classical theory in terms of oscillators and electromagnetic fields, can be elucidated in terms of a sequence of interactions between charged particles and photons.

This is most conveniently illustrated by means of space-time diagrams, known as Feynman diagrams in recognition of the physicist who first suggested their use.

For example, consider the exchange of momentum and energy that accompanies the repulsive interaction between two positively charged particles - Fig.2.26.

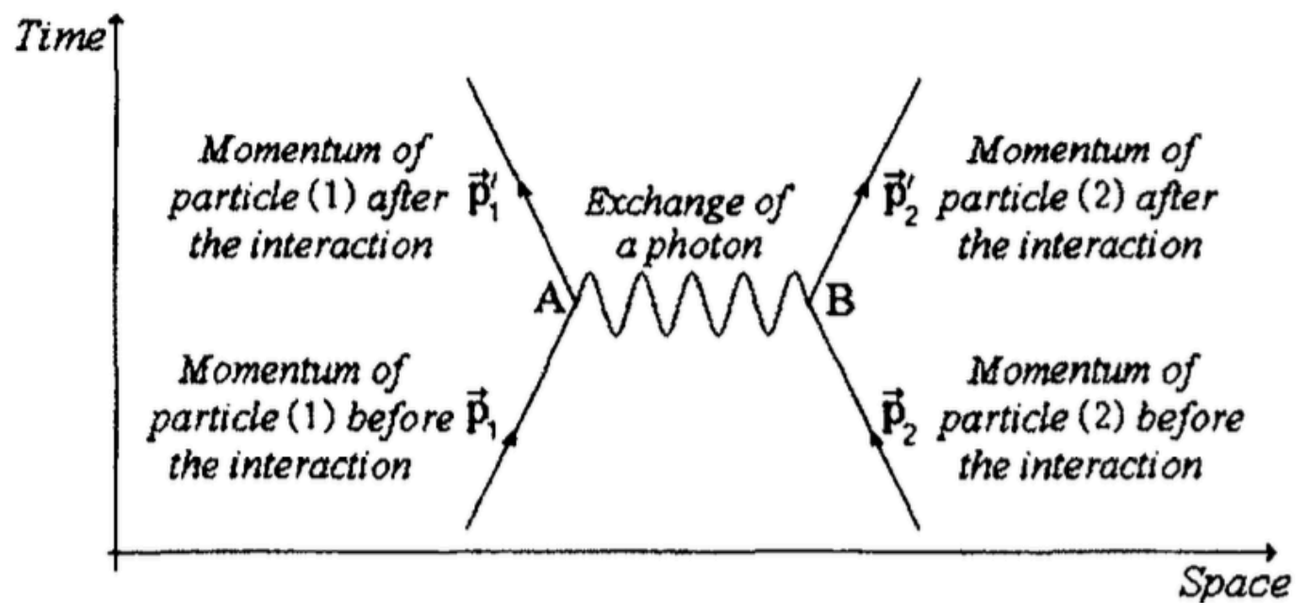


Fig.2.26 A Feynman diagram. The electromagnetic interaction between two charged particles, (1) and (2), at the moment they are situated at the points A and B, respectively, shown as an exchange of photons between them.

Suppose, that at a particular moment, one of the charged particles is at point A and the other at point B.

The first particle-particle (l)-emits a photon which moves to B where it is absorbed by a second particle-particle (2).

As a result, particle (2) absorbs energy and linear momentum and its motion is changed accordingly.

However, the motion of particle (1) must also change, for there has been a matching reduction in its energy and linear momentum; it lost the photon.

Of course, the reverse process is equally possible.

Particle (1) can absorb a photon emitted from particle (2), resulting in similar changes in their motions.

Thus, the mechanism of the interaction can be viewed as an exchange of photons between the particles.

It can be shown that the force that results from the exchange of zero rest-mass photons between two charged particles varies as the inverse square of the distance between them, i.e., in accordance with Coulomb's law

Because they are never directly observed, the photons that 'carry' the force from one particle to another, are called *virtual photons*; they appear in neither the initial nor final conditions of an experiment.

In order for the photons generated in such interactions to become actual and observable, a supply of energy must be provided.

2.3.3 Photons - Light Particles

The success of the quantum theory of radiation in elucidating the photoelectric and Compton effects reopened the debate on the nature of light.

Although the interference, diffraction and polarization of light had been convincingly explained by the wave theory, the interactions between radiation and matter appeared to offer equally convincing evidence that light is comprised of particles-photons.

So do we need both models, the wave and the particle, for a complete understanding of the phenomenon of light?

Or can the interference, diffraction and polarization of light also be explained by the particle theory of radiation?

In other words, is the quantum theory of light a partial or a complete theory of light?

We will start our investigation of this question by reference to Young's famous double-slit experiment.

As we described earlier, Young observed that a sequence of bright and dark fringes or bands -an interference pattern- appears when light from a single source is projected onto a screen through two adjacent narrow slits.

If the screen is replaced by a photographic plate, a permanent image of the interference pattern can be obtained by the usual photographic processes (Fig 2.27a).

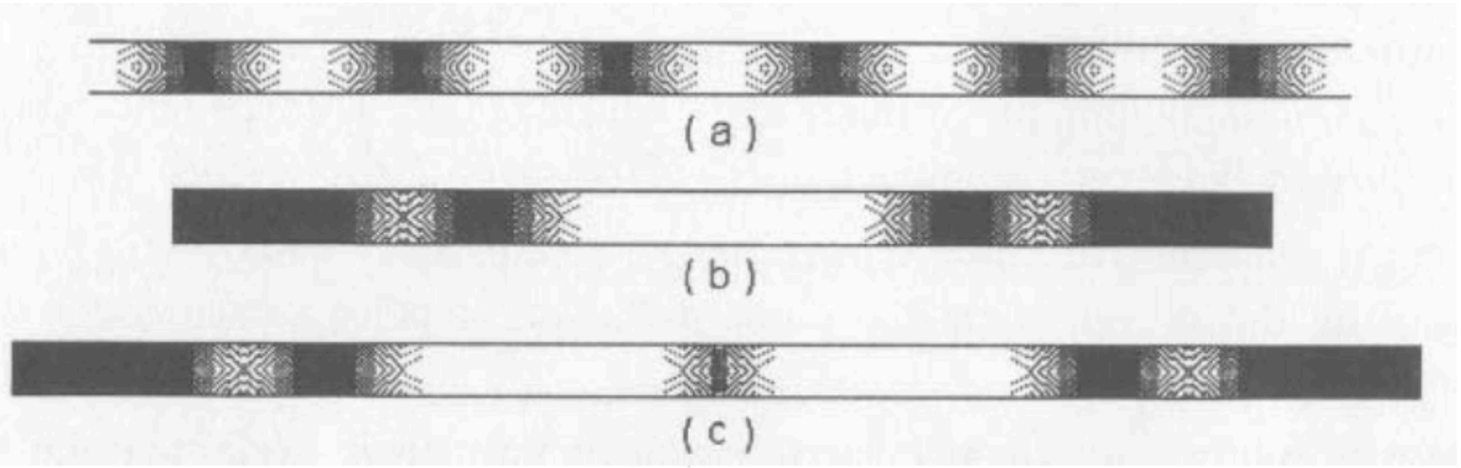


Fig. 2.27 The patterns that appear in photographs of. (a) The typical interference pattern of alternating bright and dark fringes obtained in Young's experiment when monochromatic light is projected through two adjacent narrow slits; (b) The diffraction pattern typically obtained when monochromatic light is projected through a single slit: a broad central bright band with weaker side bands; (c) The two overlapping diffraction patterns produced when each slit is opened for half the time but never both at the same time.

An interference pattern is only obtained if the light passes through both slits concurrently, i.e., if both slits are open at the same time.

If just one is open, a different pattern appears.

This pattern, characteristic of single slit diffraction, is usually referred to as a diffraction pattern (Fig. 2.27b).

If the slits are opened sequentially, first the one and then the other, but never both at the same time, the final image that appears in the photograph is of two adjacent and overlapping diffraction patterns (Fig 2.27c).

How is the formation of the interference pattern produced in Young's experiment accounted for by the wave theory of light and the quantum theory of light?

1. <u>Young's experiment according to the wave theory of light</u>.

According to Huygens' principle and the wave theory of light, the role of the two slits in the production of the interference pattern in Young's experiment is clear and well defined.

They are coherent sources of secondary light waves, i.e., they are 'orchestrated and in time with one another' such that the phase difference between them is constant.

The interference pattern results from the superposition of these secondary waves on the screen.

Bright bands-intensity maxima-appear where the secondary waves interfere constructively; dark bands-intensity minima-where they interfere destructively.

In general, the amplitude, A, of the compound wave formed by the superposition of the secondary waves at a point on the screen in Young's experiment is given by

$$A^2 = A_1^2 + A_2^2 + 2A_1A_2\cos\phi \qquad (2.48)$$

where A_1 and A_2 are the amplitudes of the secondary waves and the angle ϕ is the phase difference between them at the particular point.

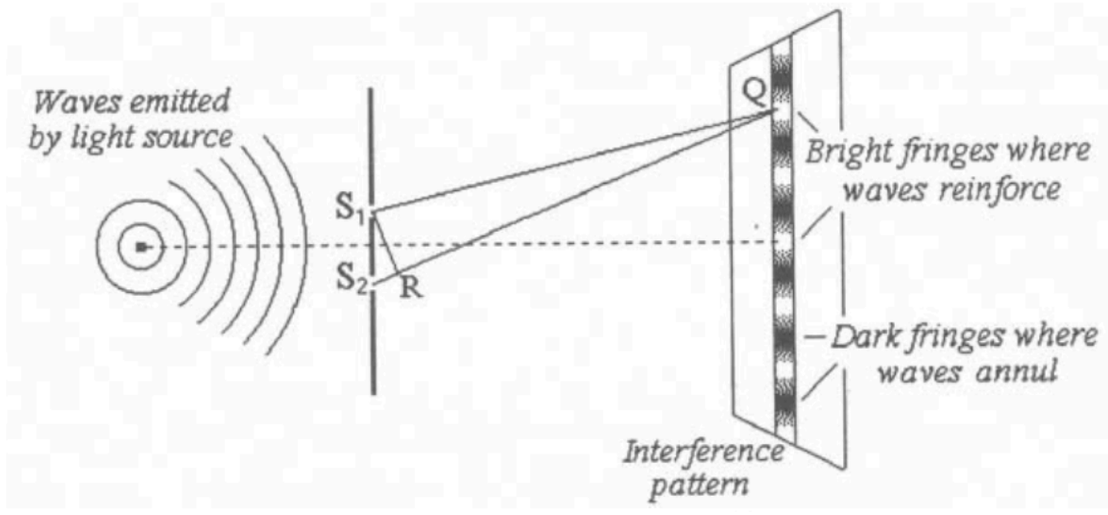
The intensity of the light at the particular point is, in turn, proportional to the square of the amplitude of the compound wave formed by the superposition of the secondary waves: $I \propto A^2$. Hence:

* **Intensity maxima (constructive interference) appear where $\cos\phi = +1$, i.e., at those places on the screen where the phase difference satisfies the condition $\phi = 2n\pi$ $(n = 0, \pm1, \pm2...)$.**
* **Intensity minima (destructive interference) appear where $\cos\phi = -1$, i.e., where $\phi = (2n+1)\pi$ $(n = 0, \pm1, \pm2...)$.**

The phase difference, ϕ, between the secondary waves arriving at a particular point on the screen depends only on the difference, δ, in the lengths of the optical paths to them point (Fig 2.28):

$$\phi = 2\pi\delta/\lambda \qquad (2.49)$$

Fig.2.28 Young's experiment. S1 and S2 are two narrow horizontal slits through which light passes from the source to the screen. The slits constitute coherent sources of secondary waves. Whether the secondary waves reinforce or annul at a particular point on the screen depends on the difference, S=S2R, between the respective lengths of the optical paths, S1Q and S2Q, from the slits to the point



where A is the wavelength.

Thus, according to the wave theory of light, the appearance of the interference pattern obtained in Young's experiment is determined by the difference in the lengths of the optical paths taken by the light from the two coherent sources.

* **Intensity maxima correspond to an optical path difference $\delta = n\lambda$.**
* **Intensity minima correspond to an optical path difference $\delta = (2n+1)\lambda/2$**

If two independent light sources are used in place of the two slits, no interference pattern is observed.

A stationary and unchanging interference pattern is only obtained if the sources are coherent, such that there is a constant phase difference between them.

TYpical sources such as flames, discharge tubes and incandescent lamps emit light intermittently, as a multitude of uncoordinated wave trains each approximately a meter long.

Consequently, two such independent and unrelated sources cannot maintain a constant phase difference between them; they cannot be coherent.

The interference patterns they produce change so rapidly, about a billion times a second, that the screen's illumination appears uniform.

However, since the sources of the secondary waves in Young's experiment -the two narrow slits- share the same original light source, they automatically and simultaneously suffer the same phase changes and so they are of necessity related and coherent.

2. Young's experiment according to the quantum theory of light.

According to quantum theory, the light intensity at a point on the screen depends on the number of photons striking there.

The darkening of an exposed photographic plate results from a chemical reaction initiated by the photons in the incident light.

Thus, the photograph of the interference pattern (and likewise that of a diffraction pattern) records the distribution of the photons striking the plate.

The bright bands in the photographic image of the interference pattern correspond to areas upon which many photons fell and the dark bands to areas upon which almost no photons fell.

However, the role of the slits in the quantum interpretation of interference is not immediately clear.

How exactly do they affect the distribution of the photons on the photographic plate? And why must both slits be kept open for an interference pattern to appear?

We will investigate these questions by considering the behavior of individual photons as they pass from the light source to the photographic plate.

In a simple experiment carried out in 1908, Taylor investigated how an interference pattern develops in Young's experiment when extremely weak light is used, so weak that the probability of there being more than one photon at any time in the space around the slits was negligible.

For this purpose, he assembled a light-proof box containing a weak light source, a series of filters, two slits and a photographic plate, arranged as required for Young's double-slit experiment (Fig. 2.29).

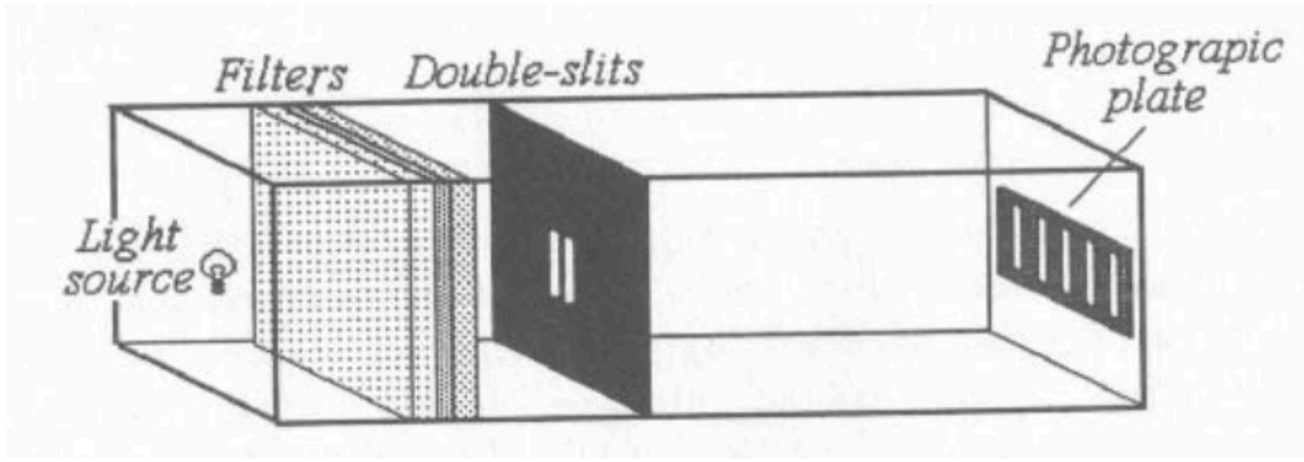


Fig. 2.29 Taylor 's Experiment. A light-proof container enclosing the components of Young's double-slit experiment. After passing through the filters, the light was so weak that the probability of there being more than one photon at any time in the space around the slits was negligible. The interference pattern appears on the photographic plate. In alternative assemblies, the photographic plate is replaced by sensitive electronic light detector that emits a signal every time a photon strikes it.

The assembled apparatus was left to stand for a few days after which the photographic plate was removed and processed.

After a fresh photographic plate had been inserted, the apparatus was again left to stand, this time for longer.

The photographic plate was then again removed and processed.

This was repeated several times, on each occasion allowing a longer time for the interference pattern to develop, up to a maximum of a few months.

When left for just a few days, no clear or distinct pattern appeared in the developed photograph; just a random sprinkling of white dots.

However, after longer periods of time, the shape and form of a typical interference pattern began to emerge in the distribution of the dots; and after still longer periods-about three months-a distinct interference pattern appeared (Fig 2.30).

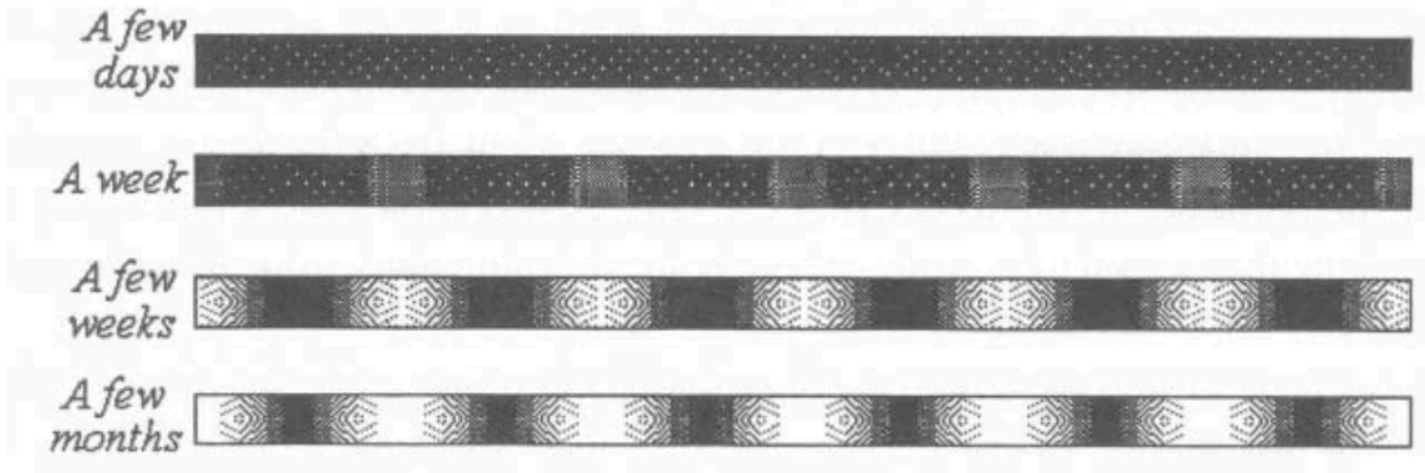


Fig.2.30 Taylor's experiment. The interference patterns that appear in the photographic images after different periods of time, from a few days (the top pattern) to a few months (the bottom one).

The appearance of the dots in the first images confirms that the interference pattern is indeed the result of the distribution of a multitude of photons on the surface of the photographic plate, each dot corresponding to the incidence of a single photon.

However, since there was never more than one photon in the space around the slits at any time, the interference pattern that emerged in the later images could not be the result of an interaction between the photons; it could not be the outcome of the constructive or destructive combination of photons.

Evidently, the interference pattern observed in Young's experiment is an inherent property of each photon; it is as though each photon 'interferes' with itself as it passes through the apparatus from the source to the screen.

Even when the light is of greater intensity and many photons are simultaneously passing through the apparatus, each interferes only with itself and not with the others.

Does this mean that each photon splits in two as it passes through the slits?

If the photographic plate is replaced by an electronic light detector that emits a signal every time a photon strikes it, or by a fluorescent screen which flashes when struck by a photon, it is found that the signals or flashes are all of the same intensity.

Similarly, if photon detectors are installed next to each slit, they never simultaneously record the passage of a photon.

There is, therefore, no evidence for a splitting of the photons during the experiment, so this cannot be the purpose for the slits.

Is it then the case that some of the photons pass through one slit and others through the other, such that when added together the joint effect of their distribution on the photographic plate is to produce the interference pattern?

To examine this possibility, the Taylor experiment is repeated but with just one slit open.

After allowing time for a pattern to develop on the photographic plate, this slit is closed and the second is then opened for the same period of time as the first.

Under these circumstances no interference pattern is produced. Instead, two adjacent overlapping single-slit diffraction pattern appear (Fig 2.27c).

Clearly, the interference pattern is not the sum of the distribution of the photons that pass through one slit plus that of the photons passing through the other.

Let us now suppose that, instead of being opened and closed one at a time, both slits are left open all the time, but a detector designed to reveal through which slit each photon actually passes is installed in the apparatus.

In this case too, two adjacent overlapping single-slit diffraction patterns are produced and not the typical double-slit interference pattern.

Moreover, it makes no difference whether the detector is placed closer to the slits than is the screen, or further away, i.e., whether the photon's choice can be detected before or after it strikes the screen; the very presence of the detector destroys the interference pattern.

If the detector is turned off or removed from the apparatus, the interference pattern reappears.

Hard as it is for us to understand, the photons appear to behave as though they were 'aware' of the intent of the experiment.

Evidently, an interference pattern only appears when the experiment provides no means of determining through which of the slits each photon passes on its way from the source to the screen, i.e., only when both slits are open concurrently and are 'unwatched'.

The patterns appearing on the screen or the photographic plate are undoubtedly the result of the distribution of the photons striking their surfaces.

However, there is no way of anticipating exactly where a particular photon will strike.

Likewise, it is impossible to tell which path a particular photon took, through one slit or the other; a photon cannot be assigned a specific trajectory.

We cannot even suppose that half the photons went through one slit and half through the other.

We can only predict how a very large number of photons will be distributed on the surface of the screen or photographic plate.

On this view, the interference and diffraction patterns are statistical phenomena, associated with the probability that, in a given experimental set-up, a photon will strike a particular point on the screen.

The greater the probability, the more photons strike the point and the stronger the light intensity at the point.

Hence, the bright fringes correspond to a high probability; the dark fringes to a low probability.

The essential features of the explanations given by the wave and quantum theories of light to Young's experiment are summarized in the following matrix:

	Wave Theory	Quantum Theory
Pattern of fringes:	The intensity bands arise from the variations in the value of the square of the amplitude, A^2, of the resultant wave at each point on the screen.	The intensity bands arise from the variations in the probability, P, of a photon striking different points on the screen
Role of the slits:	To act as two coherent sources of the secondary waves that interfere on the screen.	To present two potential routes by which a photon can pass from the source to the screen

Is light then a wave or a particle phenomenon?

Can the two explanations, that of the wave theory in terms of wave amplitudes and that of the photon theory in terms of counting photons, be reconciled?

Or are the wave and particle hypotheses of light in fact as irreconcilable as classical physics had taught?

To answer these questions, let us reconsider the experiment conducted by Compton on the scattering of X-rays.

In order to measure the wavelength of the X-rays, Compton employed the technique developed by Bragg which imputes a wave character to the X-rays.

However, the theory and analysis upon which Compton predicted the change he expected to find in the wavelength of the radiation was based on the assumption that X-rays are composed of photons. i.e., particles, that collide with free electrons according to the laws of mechanics.

Thus, overall, we may say that Compton's experiment demonstrated that X-rays have both particle and wave character, each attribute manifesting itself independently in a different part of the experiment in line with the experimenter's expectations.

In modern physics, the wave and particle hypotheses of light complement one another; they are not mutually exclusive.

In our understanding of light intensity we link these two qualities by saying that the number of photons, N, striking a particular point on a photographic plate is directly proportional to the intensity, I, of the light wave formed at the point, i.e., to the square, A^2, of its amplitude:

$$N \propto A^2 \tag{2.50}$$

Thus, in Young's experiment, the probability, P, of a photon striking close to a particular point on the screen will be directly proportional to the square of the amplitude of the compound wave formed at the point by the superposition of the secondary waves:

$$P \propto A^2 = A_1^2 + A_2^2 + 2A_1A_2 \cos\phi \tag{2.51}$$

At those points where the interference is constructive, the square of the amplitude is a maximum and so the probability of a photon striking the point is high.

On the other hand, at those points where the interference is destructive the amplitude is a minimum and the probability of a photon striking the point is negligible.

The single-slit diffraction pattern is also a statistical phenomenon; it too reflects the probability of a photon striking different points on the photographic plate.

If just one slit is open, the probability, P_1, that a photon will strike a particular point depends on the square of the amplitude, A^2_1 at the particular point of the secondary waves it emits.

Overall, this probability corresponds to a single-slit diffraction pattern.

Likewise, when just the other slit is open, the probability, P_2, that a photon will strike any point depends on A^2_2.

Overall, this probability corresponds to a single-slit diffraction pattern too.

When the slits are opened one at a time, one after the other, the total probability, P_{12}, that a photon will strike a particular point on the photographic plate is simply the sum of the two separate probabilities:

$$P_{12} = P_1 + P_2$$

or in terms of the wave amplitudes:

$$P_{12} \propto A_1^2 + A_2^2$$

This probability, which corresponds to two single-slit diffraction patterns, is different from the probability,

$$P \propto A_1^2 + A_2^2 + 2A_1A_2 \cos\phi\,,$$

associated with the formation of an interference pattern; the interference factor,

$$2A_1A_2 \cos\phi\,,$$

is missing.

Installing a photon detector to determine through which slit each photon passed is removes any uncertainty we may have as to the path taken by the photons.

As such, it is essentially equivalent to opening the slits one at a time and so it produces the same distribution pattern.

2.3.4 The Locality Paradox

When first proposed, the elusiveness attributed to photons by the quantum theory of light troubled many physicists.

Known as the Copenhagen interpretation, by virtue of its advocacy by the Danish physicist Bohr, this construal of quantum theory was vigorously opposed by no less a figure than Einstein himself. “God does not play dice”, he declared.

If photons actually exist, their motion should be determinable and not ‘spooky’ as the Copenhagen interpretation asserted.

In 1935, Einstein proposed an experiment based on the polarization of light, that he believed would demonstrate the inadequacy of the Copenhagen interpretation.

The quantum theory of light accounts for light polarization by assigning an intrinsic angular momentum -spin- to the photons.

In principle, photon spin can take two forms: it can either be parallel or anti-parallel to the direction of the photon’s linear momentum.

The variety of modes and orientations of the polarization of light can all be accounted for in terms of these two basic polarization states.

Suppose, Einstein argued, that a stationary particle emits two photons that move off, back-to-back, in opposite directions.

The conservation of angular momentum requires that the photons' spins be in opposite orientations, such that their sum be zero.

Thus if the spin of one is subsequently measured, that of the other will automatically become known too.

In the context of the realism that underlies classical physics, this 100% correlation between the spins presents no difficulties.

It asserts that the particles actually possessed their respective spins from the moment they were emitted; the measurement just reveals this fact.

However, the Copenhagen interpretation of the quantum theory of light sees things differently.

It asserts that, each photon has an equal probability of being in the parallel or anti-parallel spin orientation.

A photon does not 'know in advance' how it will be spinning when the measurement is made, just as we cannot tell through which slit a photon passes in Young's experiment; each orientation or pathway has the same probability.

Accordingly, we cannot tell whether a particular photon was emitted with its spin in one orientation or another.

All we can say is that each photon has a 50% chance of being polarized in one or other of the two possible orientations.

And so things remain, until the spin of one of the photons is actually measured.

This act instantly fixes the spin orientations of both photons.

For if the measured photon is found to have one orientation, the other photon must, of necessity, have the opposite orientation.

Should its spin be subsequently measured, it will be found to be in the opposite orientation to that found for the first photon.

But what if the two photons had travelled some distance from one another when the first measurement was made.

How does the second photon 'learn' that this measurement has been made and that the orientation of its spin has, of necessity, been fixed from that very moment?

What mechanism ensures that the second measurement accords with the first, as required by the laws of conservation?

The controversy surrounding this issue became known as the *locality* paradox.

Einstein concluded that these difficulties could only be reconciled in one of the following two ways:

1. **If, at the very moment the first measurement is made, the result communicates itself instantaneously to the second particle and sets its spin state accordingly.**
2. **The quantum theory of light is incomplete and does not fully specify the actual 'goings-on'; there are *hidden variables* which find no expression in conventional quantum theory but without which a full and *real* explanation cannot be given.**

The first alternative clearly violates special relativity, under which the instantaneous transmission of information is prohibited.

Not surprisingly, it was not Einstein's choice.

He opted instead for the hidden variables.

However, subsequent studies have shown that they too exhibit a non-locality every bit as paradoxical as that he sought to avoid.

It transpires, that "any realist interpretation of quantum theory in which it is deemed meaningful to say that an individual system possesses the values of its physical quantities in a way that is analogous to classical physics will exhibit non-locality."

A number of experiments of the type suggested by Einstein have been carried out in recent years.

Their design ruled out the possibility of any link between the production of the photons and the measurements made upon them or that the detectors could have affected the findings.

They all produced results consistent with the Copenhagen interpretation.

The overwhelming conclusion, seventy years after it was first put forward, is that the ‘spooky’ predictions of the Copenhagen interpretation are vindicated.

Accordingly, it is this pragmatic construction of quantum theory that we will adopt throughout this discussion.

Part 2.4 The Mechanics of Minute Particles

The interpretation given by the quantum theory of radiation to the interference and diffraction of light rests on the principle that an explicit trajectory cannot be assigned to the motion of an individual photon.

We can only state that the photon was at point A at one moment and appeared at point B some time later.

Along which of the potential routes linking A and B it actually went can never be known.

All that we can ever determine is the probability, P, that it took a particular path.

Furthermore, we saw that this is equivalent to ascribing a wave amplitude, A, to each potential path whose square, A^2, is proportional to the probability that the photon took that path.

We now investigate whether this elusiveness is a feature only of photons or is a general property of all minute particles.

2.4.1 De Broglie's Hypothesis

According to the evidence of our senses, material objects have well defined shapes and sizes, i.e., they are localized in space.

However, the theory of special relativity asserts that the most distinctive attribute of material objects -their mass, is essentially equivalent to the most distinctive attribute of the fields associated with the interactions between them -their energy.

The theory states that the two -mass and energy- are interchangeable forms of the same fundamental physical quantity.

Hence, we might reasonably expect that what pertains to the one should pertain to the other.

This is the essence of the hypothesis put forward by De Broglie in his doctoral thesis in 1924:

> ***Just as the energy and momentum associated with electromagnetic waves can be interpreted in terms of particles - photons, the dynamical properties of material particles, such as their energy and momentum, can be interpreted in terms of waves - 'matter waves'.***

Quantum theory interprets electromagnetic fields in terms of particles -photons.

To be consistent, we should be able, in the reverse manner, to interpret material particles in terms of fields -matter fields.

De Broglie proposed, that the two fundamental mathematical formulae, E = hf and p =h//λ previously thought to bear only on photons, were applicable to all particles, whatever their nature.

The inference is that every material particle has a matter wave associated with it of frequency f = E/h, where E is the total energy of the particle.

The wavelength, λ, of the particle's matter wave is given by:

$$\lambda = \frac{h}{p} = \frac{h}{mu} \tag{2.52}$$

where p = mu is the particle's relativistic linear momentum. Equation (2.52) is known as *De Broglie 's Equation*.

The import of De Broglie's hypothesis is that every particle flow can be viewed as a propagating disturbance in a field, i.e., as a wave motion.

Just as light, that was traditionally thought of as a wave, can be understood as a stream of particles -photons, so any stream of material particles can be understood as a wave.

De Broglie's revolutionary hypothesis was confirmed in 1927 when Davisson and Germer in the United States and Thomson in the United Kingdom independently discovered that electrons can be diffracted.

The American team was studying the scattering of low energy electrons from metal surfaces.

The surfaces had to be scrupulously clean.

When a narrow beam of electrons was aimed at a nickel crystal that had been heated in hydrogen to remove any trace of oxides from its surface, exceptionally high intensities of scattered electrons were found at specific angles relative to the incident beam (Fig 2.31).

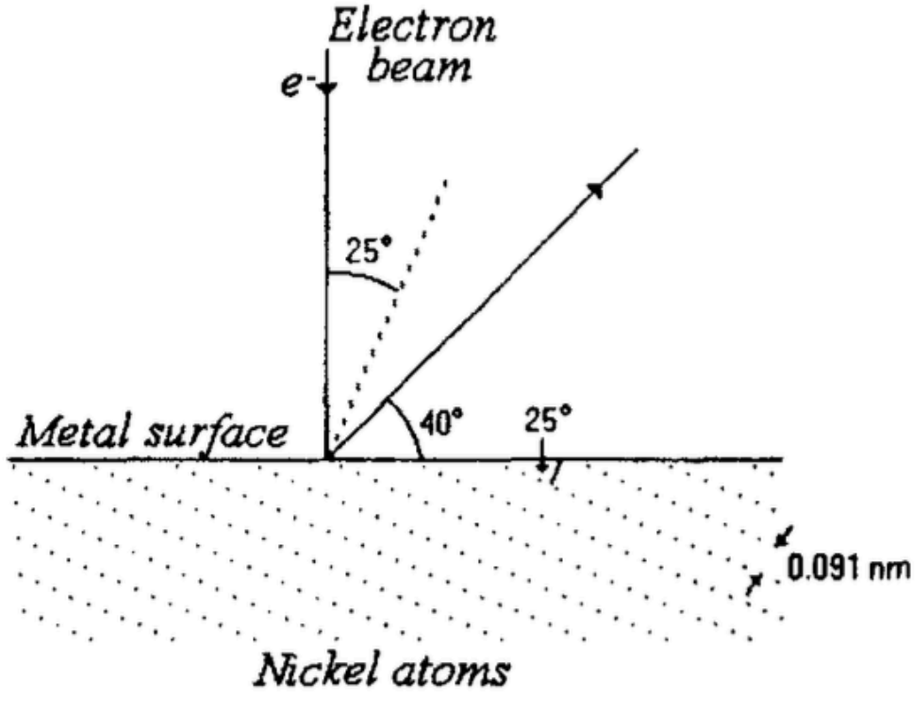


Fig 2.31 The Davisson-Germer experiment: the scattering of electrons from a nickel surface. Electrons accelerated through a potential difference of 54V, were strongly scattered at an angle of 50° to the incident beam. The angle between the reflected beam and the plane of the atoms is 65°.

In terms of waves, this experiment is similar to that in which the diffraction of X-rays by crystal planes was first demonstrated by Bragg, except that in this case it is the matter wave associated with the electrons that is diffracted and not a high frequency electromagnetic wave.

The electrons are reflected at the atomic planes in the metal crystal, and their reflected matter waves interfere constructively under the conditions for constructive interference given by the Bragg equation:

$$n\lambda_{electron} = d\sin\theta_n \tag{2.53}$$

At the time the experiment was conducted, the spacing between the atomic planes in nickel crystals had already been determined from X-ray diffraction experiments and was known to be d = 0.091nm.

Substituting this in the Bragg equation gives for the wavelength of the electron's matter wave (n = 1):

$$\lambda = 2d\sin\theta = 2\cdot 0.091\cdot\sin 65^\circ = 0.165\text{nm} \tag{2.54}$$

As the following calculation shows, this wavelength almost exactly equalled that predicted by the De Broglie hypothesis for the matter waves associated with electrons accelerated from rest through a potential difference of 54V.

The linear momentum of a particle, in terms of its mass, m, and its kinetic energy, K, is given by

$$p = \sqrt{2mK} \tag{2.103}$$

Substitution in the De Broglie equation gives:

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2mK}} \tag{2.68}$$

In general, the kinetic energy of the electrons accelerated from rest by a potential difference V is $K = eV$. This gives for the wavelength of the matter wave of an electron accelerated from rest by a potential of 54V (a 'slow' electron whose velocity is about $4.3 \cdot 10^6$ m/s):

$$\lambda = \frac{h}{\sqrt{2meV}} = \frac{6.63 \cdot 10^{-34}}{\sqrt{2 \cdot 9.11 \cdot 10^{-31} \cdot 54 \cdot 1.6 \cdot 10^{-19}}} = 0.167\text{nm} \tag{2.69}$$

The close correspondence between the experimental value and the value computed from De Broglie's equation, was seen as confirmation of the hypothesis

The diffraction of other minute particles, such as neutrons, protons and atomic nuclei, was subsequently demonstrated in other experiments.

In each case, the measured wavelengths of the matter waves corresponded to those predicted by De Broglie's hypothesis.

Furthermore, interference patterns of alternating intensity maxima and minima were obtained in double-slit experiments, similar to that carried out by Young, but using an electron source in place of the light source.

These experiments established the wave character of material particles beyond any doubt.

Moreover, just as in the case of photons, individual material particles defied all attempts to determine precisely how they moved from one point to another.

It proved as impossible to assign a definite trajectory to individual material particles as it had been for photons.

Elusiveness is a property of all minute particles.

The elusiveness and particle-wave complementarity exhibited by minute particles is a difficult concept to accept.

Born has suggested the following elucidation:

The ultimate origin of the difficulty lies in the fact (or philosophical principle) that we are compelled to use the words of common language when we wish to describe a phenomenon....Classical physics has restricted itself to the use of concepts of this kind; by analyzing visible motions it has developed two ways of representing them by elementary processes: moving particles and waves. There is no other way of giving a pictorial description of motions-we have to apply it even in the region of atomic processes, where classical physics breaks down. Every process can be interpreted either in terms of corpuscles or in terms of waves, but on the other hand it is beyond our power to produce proof that it is actually corpuscles or waves with which we are dealing, for we cannot simultaneously determine all the other properties which are distinctive of a corpuscle or a wave, as the case may be. We can therefore say that the wave and corpuscular descriptions are only to be regarded as complementary ways of viewing one and the same objective process...The corpuscular description means at bottom that we carry out measurements with the object of getting exact information about momentum and energy relations (e.g. in the Compton effect), while experiments which amount to determinations of place and time we can always picture to ourselves in terms of the wave representation (e.g. passage of electrons through thin foils and observations of the deflected beam).

In general, the wave nature of a phenomenon is only evinced when the dimensions of the obstacles or apertures that the wave encounters are no more than just a few fold larger than its wavelength.

The wavelength of electrons with energies of a few hundred electron-volts, such as those which generate the images on oscilloscope and television screens, are of the order of magnitude of an atom, 10^{-10}m,

Consequently, when considering the motion of such electrons through oscilloscopes or television tubes — systems of macroscopic proportions— we can ignore the electrons' wave character and describe them as classical particles.

However, when describing the motion of the electrons in an atom, their wave character cannot be ignored; indeed, without taking it into consideration, many of the properties of the atom defy comprehension, as we will see this later.

The wavelengths of the matter waves associated with macroscopic particles are so small that their wave properties are never discerned.

Thus, for example, the De Broglie wavelength of a grain of sand of mass 10^{-6}kg (one milligram) moving with a velocity of l0 m/s, is:

$$\lambda = \frac{h}{mu} = \frac{6.63 \cdot 10^{-34}}{10^{-6} \cdot 10} = 6.63 \cdot 10^{-29}\,\mathrm{m} \qquad (2.55)$$

Obstacles or apertures of this order of magnitude do not exist and so under normal conditions the motion of macroscopic particles can be described by classical Newtonian mechanics.

The resolving power of a microscope depends on the wavelength of the waves with which it 'sees'; the smaller the wavelength the greater its power of resolution.

A light microscope can reach a maximum useful magnification of about x500.

Since electrons carry an electric charge, their motion is affected by electromagnetic fields; a Lorentz force acts on them.

This makes it possible to focus the electrons in the electron microscope by means of magnetic lenses.

Electrons accelerated through a potential of 100,000V have a wavelength of 0.0037nm, which is about 100,000 smaller than that of visible light.

Consequently, a microscope operating with the matter waves associated with such electrons, should have a power of resolution many times greater than that of a light microscope.

Such a microscope is called an *electron microscope*.

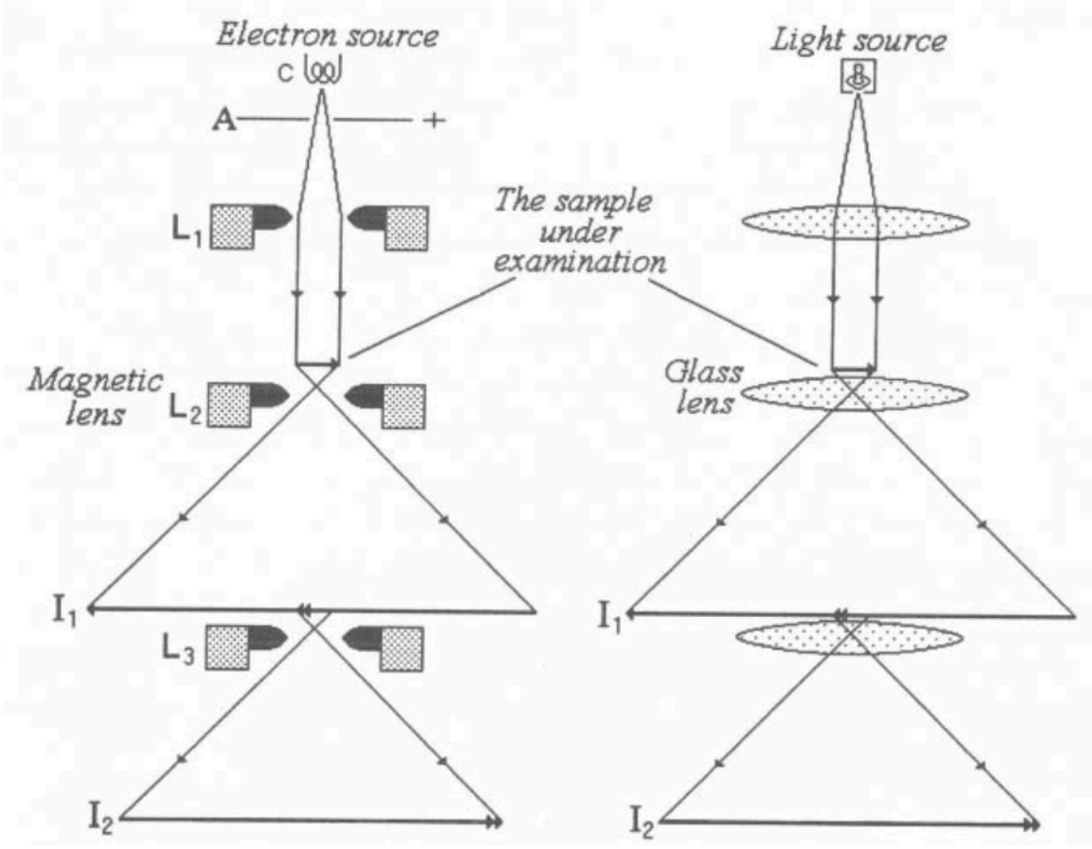
In practice, the useful magnification obtained is less than the theoretical because of limits set by the magnets.

Nevertheless, magnifications of x1,000,000 can be achieved, making possible the examination of the most minute constituents of a system.

Two types of electron microscopes have been developed: transmission and scanning.

The design of the transmission electron microscope closely resembles that of an optical microscope (Fig 2.32).

Fig 2.34 A light microscope and a transmission electron microscope. The components of an electron microscope match those of the light microscopes. The sample being examined is irradiated with electrons instead of with light. The electrons are accelerated from the cathode C to the anode A. The magnetic lens L_1 collimates the beam incident on the sample. The lens L_2 produces the first image I_1, and the lens L_3 projects the final image I_2 on a photographic plate or television screen.



To avoid excessive attenuation of the electron beam the sample being examined must be sliced very thin.

As a result, only two-dimensional images are obtained with the transmission microscope.

The scanning electron microscope produces a seemingly three-dimensional image of the sample being examined.

The samples magnified in these microscopes must be electrically conducting.

Samples of nonconducting material such as biological specimens have a thin metal layer evaporated onto their surface.

The electron beam is focused to a fine spot (-l0nm) that scans the surface of the sample releasing secondary electrons from it.

The number of secondary electrons depends on the nature of the surface and the image is formed by collecting these secondary electrons.

A more recent development is a scanning electron microscope that employs the tunnel effect.

In this device the tip of a very fine pointed metal probe moves over and across the surface of the conducting sample.

A potential difference is maintained between the probe and the sample's surface.

Although the tip does not actually touch the surface, electrons can tunnel across the narrow gap between them.

By plotting the tip-surface current or the movement of the tip required to maintain a constant current, the sample's surface can be mapped.

Resolutions of less than a tenth of a nano-meter are routinely achieved with this microscope.

2.4.2 Heisenberg's Uncertainty Principle

Both classical and relativistic mechanics assert that, in principle, there is no constraint on what can be known about the physical parameters of a system; any uncertainty in the actual value of these quantities results only from human frailty or ignorance.

Thus, one of the postulates of classical mechanics is that the position and linear momentum of a particle can be exactly determined at any moment in time.

A sequence of such determinations, defines the particle's trajectory through space-time.

Furthermore, classical and relativistic mechanics are fundamentally deterministic, i.e., they assert that from a knowledge of the state of a mechanical system at any moment and of the laws governing it, everything about its past and its future can be ascertained.

In practice, exact predictions may be hard or even impossible to achieve due to practical or technical inadequacies in the accuracy of our measurements, but the possibility of perfectly accurate measurements is not a priori precluded.

The confidence expressed in these assertions mirrors the 18th century European belief that nature is reliable and that there is no natural limit to what man can know about the world; that all knowledge is just 'out there' waiting to be acquired.

Nothing is in principle inexplicable; uncertainty is just a feebleness of the mind.

Given time, all the mechanisms that govern the universe would be discovered.

The great expectations of this mechanistic philosophy are illustrated by the following extract from the writings of the great French physicist Laplace.

> *All events, even those which on account of their insignificance do not seem to follow the great laws of nature, are a result of it just as necessarily as the revolutions of the sun. In ignorance of the ties which unite such events to the entire system of the universe, they have been made to depend upon final causes or upon hazard (chance] ... But these imaginary causes have gradually receded with the widening bounds of knowledge and disappear entirely before sound philosophy, which sees in them only the expression of our ignorance of the true causes. Given for one instant an intelligence which could comprehend all the foxes by which nature is animated and the respective situation of the beings who compose it - an intelligence sufficiently vast to submit these data to analysis it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atoms; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers…a feeble idea of this intelligence. The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance.*

In the deterministic view of Laplace, we could even predict how a dice would fall, were it not for our benightedness.

Precise information as to its starting position, together with a knowledge of the laws of motion, is all that we would need.

But the inherent elusiveness of those minute particles that manifest a wave character raises the question of whether, a priori, their position and linear momentum can be simultaneously determined.

In principle, is such an observation possible?

We will examine this question by means of the following thought experiment.

The power of resolution of a microscope depends on the wavelength, λ, of the waves with which it 'sees'.

Because of diffraction effects, well defined images of objects, whose size is comparable to the wavelength, cannot be obtained; the position of such objects cannot be exactly determined.

The uncertainty, Δx, in the object's x co-ordinate, is of the order of magnitude of the wavelength:

$$\Delta x \approx \lambda \tag{2.56}$$

the shorter the wavelength employed, the greater the linear momentum, p , associated with the waves and the greater will be the concomitant change in the object's motion.

In the collision between the waves and the object, the maximum possible change, Δp_x, in the object's linear momentum parallel to the X-axis will be of the order of magnitude of the linear momentum of the incident waves, i.e.,

$$\Delta p_x \approx h/\lambda \tag{2.57}$$

Putting all these results together gives

$$\Delta x \cdot \Delta p_x \approx h \tag{2.58}$$

The significance of this relationship is that the smaller the uncertainty, Δx, in the object's position the greater the uncertainty, Δp_x, in its linear momentum, and vice versa.

Employing shorter wavelengths would indeed improve the measurement of the object's position, but this would be achieved at the expense of a loss of accuracy in the measurement of its momentum.

Hence, in principle, exact simultaneous measurements of an object's position and linear momentum cannot be made.

A rigorous mathematical analysis shows, that the actual relationships between the inherent uncertainties in a particle's space co-ordinates and the components of its linear momentum are

$$\begin{aligned} \Delta x \cdot \Delta p_x &\geq h/4\pi \\ \Delta y \cdot \Delta p_y &\geq h/4\pi \\ \Delta z \cdot \Delta p_z &\geq h/4\pi \end{aligned} \tag{2.59}$$

Equation (2.58) is clearly consistent with these relationship.

This limitation on the simultaneous measurement of complementary physical quantities is known as *Heisenberg 's Uncertainty Principle*.

The quantity h/2π which occurs frequently in quantum physics, is usually designated by the symbol $\hbar$ called "h-bar".

The uncertainty principle can also be expressed in terms of the inherent uncertainty, ΔE, in a particle's energy and the inherent uncertainty, Δt, in a the characteristic evolution time interval:

$$\Delta E \cdot \Delta t \approx \hbar \tag{2.60}$$

For example, this relationship relates the uncertainty, ΔE, in the energy of the photon emitted when an atom falls from an excited stationary state to the ground state to the uncertainty, Δt, in the lifetime of the excited state.

The uncertainty in the lifetime of the excited state is of the order of its lifetime.

It follows, that the more stable the state the greater is Δt and the smaller is ΔE.

In general, if the energy of an excited stationary state is accurately determined, as it usually can be by spectroscopic measurements of the frequency of the photons emitted in transitions from it, no accurate information can be obtained as to the moment the individual photons were actually emitted.

Similarly, a particle can 'borrow' an amount of energy, ΔE, from apparently empty space without violating energy conservation, so long as it returns the loan within a period of time $\Delta t \approx \hbar/\Delta E$.

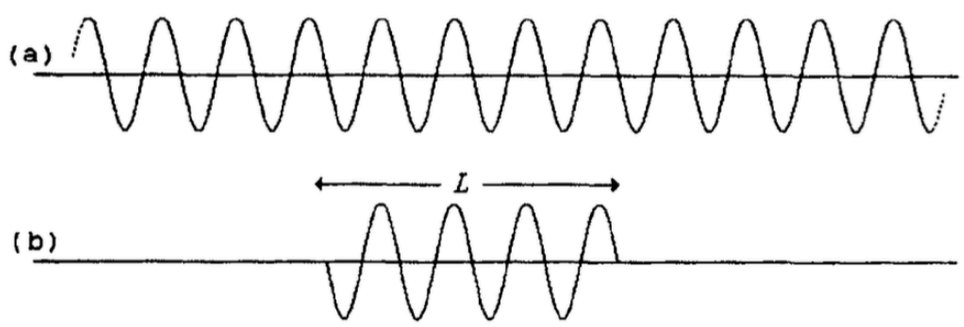
The creation, apparently *ex nihilo*, of the virtual photons which mediate the rapid interactions between particles described by Feynman diagrams, is explained in this way.

We emphasize that the uncertainties expressed in the expressions such as $\Delta x \Delta p_x \approx \hbar$ and $\Delta E \Delta t \approx \hbar$ are not the result of technical or practical difficulties in making the measurements; they are inherent in the wave character of material objects.

They are a consequence of the fact that every wave phenomenon has some U priori uncertainty associated with it.

For example, measuring the frequency of a musical note requires a wave train of certain length (Fig 2.36).

Fig 2.36 (a) A section of an infinite sinusoidal wave. The frequency or wavelength can be measured with almost complete accuracy if a sufficiently long section is taken. (b) A finite wave train of length L. Although the frequency and wavelength can be measured with considerable accuracy, there is always an uncertainty as to exactly where the wave train starts and ends.



In order to obtain the frequency, we divide the length of time that the note is heard by the number of periods counted during that time, i.e., by the number of periods in the wave train.

The longer the note, the more accurate will be the measurement of its frequency.

On the other hand, limiting the time allowed for the measurement reduces its accuracy.

In the extreme case, if the wave train is less than one period long, the measurement of the note's frequency becomes meaningless.

Ascribing a precise trajectory to a moving particle implies absolute certainty as to its position at any moment.

Such trajectories are typically delineated in classical and relativistic mechanics by continuous functions such as E(x,t) or p(x,t), that describe the energy or linear momentum of a particle moving parallel to the x-axis in terms of its space and time co-ordinates.

However, the uncertainty principle precludes the validity of such functions *a priori*.

Why then do classical and relativistic mechanics work so well?

The reason is the diminutive size of Planck's constant, h.

The fundamental uncertainly prescribed by Heisenberg's uncertainty principle in a macroscopic object's position, $\Delta x \approx \hbar/\Delta p_x$ is totally negligible compared to its dimensions.

The motion of macroscopic objects can be properly described by the deterministic mathematical functions of classical and relativistic mechanics because the inherent uncertainties associated with them are too small to be detectable.

This is a further example of the correspondence principle.

However, the fundamental uncertainties in the position or momentum of minute particles are not negligible relative to their dimensions and so their motion cannot be described by deterministic mathematical functions.

Consequently, when setting out to examine the behavior of minute particles, we must first decide which physical quantity is the more important to us: its position or its momentum; its energy or the stability of its present state.

Information about the one is always acquired at the expense of information about the other.

Precise knowledge of a particle's momentum will leave us with little more than probabilities about its position.

Similarly, accurate measurement of the particle's energy in a particular state leaves us ignorant of how long the particle remains in the state.

Any particle property or physical variable that can be measured is termed an observable.

In classical and relativistic physics, the measurement of observables is not subject to any inherent uncertainty, only to technical or random errors.

However, in quantum physics, the degree to which a physical variable can be precisely measured is subject to the uncertainty principle.

Repeated errorless measurements of the variable will not always give precisely the same value.

The spread (variance) in the errorless measurements of a variable reflects the inherent uncertainty in the particular particle property.

If the spread is small, the variable is said to be a sharp variable; if it is large, it is said to be a fuzzy variable.

The values of a quantum system's sharp variables are called their characteristic, proper or eigenvalues.

With a few exceptions, the sharp observables are the same quantities as those conserved in classical physics.

2.4.3 Matter Waves

The concept of a matter wave defies simple intuitive visualization.

Matter waves do not correspond to vibrations in an underlying physical medium nor are they the result of particle vibrations.

A matter wave is the moving particle itself, conceived of as a disturbance in a field-the matter field.

Unlike mechanical, electrical and magnetic fields, matter fields can be neither detected nor mapped experimentally,

The quantities that oscillate in sound and electromagnetic waves, respectively, the air pressure and the magnitude of the electric and magnetic fields, have a clear physical reality and can be measured and interpreted experimentally.

The former can be measured with a pressure gauge, the latter with a test charge.

However, the analogous quantity-the 'thing'-that oscillates in matter waves can neither be measured nor interpreted in terms of any experiment.

The ‘thing’ that oscillates in matter waves is usually referred to as the Wave Function or the Probability Amplitude.

Wave functions (or probability amplitudes) are usually designated by the Greek letter psi, Ψ .

In general, the wave function, Ψ , is a complex number whose value is a function of the space coordinates x, y and z and of the time co-ordinate t.

That part of the wave function that depends only on the space co-ordinates is known as the time-independent wave function and is usually designated by the lower case letter ψ.

Although no direct physical meaning can be ascribed to the wave function, matter waves exhibit all the typical wave properties.

Thus, for example, the superposition of two or more wavefunctions produces another wavefunction.

Moreover, although it has no direct physical meaning itself, the wave function contains all the information that can be known about the particle.

This information can be extracted from the wave function by performing certain mathematical operations on it and can be used to make calculations and predictions about the outcome of experiments.

Thus, for example, the square of the absolute value of the wave function, $|\Psi|^2$, is a measurable quantity and can be interpreted experimentally.

It represents the intensity of the matter wave and indicates the probability, P, of finding the particle around a given point:

$$P \propto |\Psi|^2 \tag{2.61}$$

$|\Psi|^2$ is called the probability density.

This interpretation of the square of the wave function was first proposed by Born in 1926.

It can be drawn, by analogy, from the link between the square, A^2, of the amplitude of an electromagnetic wave at a given point (its intensity at the point) and the probability, P, of finding a photon there.

A corollary of this interpretation is that, in the one dimensional case, the probability, P(x)dx ,of a particle being found between x and x + dx at time t is

$$P(x)dx = |\Psi(x,t)|^2 dx \tag{2.62}$$

and the probability, P_{ab}, of finding the particle between x = a and x = b is

$$P_{ab} = \int_a^b |\Psi(x,t)|^2 dx \tag{2.63}$$

Clearly, the particle must be somewhere and so the sum of the probabilities for all the possible values of x must equal unity:

$$\int_{-\infty}^{\infty} |\Psi(x,t)|^2 dx = 1 \tag{2.64}$$

Any wave function satisfying equation (2.64) is said to be normalized.

Although repeated measurements of a physical variable may vary because of its inherent uncertainty, an average or expectation value can always be assigned to it.

For example, repeated measurements of a particle's x co-ordinate may reveal it to be in a different position each time.

However, if we know the particle's wave function, we can find the expectation (average) value, $\langle x \rangle$, of its position from the equation

$$\langle x \rangle = \int_{-\infty}^{\infty} xP(x)dx = \int_{-\infty}^{\infty} x|\Psi(x,t)|^2 dx \qquad (2.65)$$

In the same way, expectation values can be found for all the particle's other physical variables.

For example the expectation value of its potential energy, U, is given by

$$\langle U \rangle = \int_{-\infty}^{\infty} U(x)P(x)dx = \int_{-\infty}^{\infty} U(x)|\Psi(x,t)|^2 dx \qquad (2.65a)$$

The formal condition that a physical variable must fulfill if its values to be sharply observable is that the effect upon the wave function of a mathematical operation specific to the variable, e.g., multiplication by a given function or taking a certain derivative, must be simply to return a multiple of the wave function. In mathematical notation this is expressed as

$$[E]\Psi = E_n\Psi$$

where E in the particular variable, [E] is the specific mathematical operation and E_n is a value of the variable.

It transpires that this is only possible for certain wave functions called the eigenfunctions and only for certain values of the variable -its eigenvalues.

2.4.4 Wave Functions and Probability Amplitudes

Wave functions can also be used to describe real events such as the displacement of a particle from one point to another.

This is the basis of the new mechanics -Wave Mechanics - developed by Born, Heisenberg and Shroedinger during 1925 and 1926.

In place of the equations of motion of classical mechanics, from which the particle's exact position in space at every moment can be computed, this new mechanics provides wave functions that contain all that can be known about the particle subject to the uncertainty principle.

In principle, if not always in practice, the interpretation of actual events in terms of wave functions is a simple technique.

It involves just the following three rules.

Rule 1

When an event, all of whose initial and final conditions have been fully specified, can take place in a number of alternate ways, it being impossible to determine which one was actually taken, the wave function for the event, Ψ, is given by the sum of the wave functions, Ψ_1, Ψ_2, Ψ_3..., that represent, respectively, each of the different alternative ways:

$$\Psi = \Psi_1 + \Psi_2 + \Psi_3 + \ldots \tag{2.66}$$

This superposition of the wave functions results in interference, and the overall probability of the event, P, is given by

$$P \propto |\Psi|^2 = |\Psi_1 + \Psi_2 + \Psi_3 + \ldots|^2 \tag{2.67}$$

Constructive interference corresponds to a high probability of the event's occurrence; destructive interference to a low probability.

Rule 2

If the way the event took place can be established, there is no superposition of the wave functions. In this case, the overall probability of the event, $P_{123\ldots}$, is given by the sum of the separate probabilities for each alternative way:

$$P_{123\ldots} = P_1 + P_2 + P_3 \ldots \propto |\Psi_1|^2 + |\Psi_2|^2 + |\Psi_3|^2 + \ldots \tag{2.68}$$

Rule 3

When the event can only take place in a sequence of stages, one after the other, its wave function is obtained from the product of the wave functions of each stage:

$$\Psi = \Psi_1 \Psi_2 \Psi_3 \ldots \tag{2.69}$$

In order to demonstrate how wavefunctions are used to describe real physical phenomena, we will consider a double-slit experiment similar to Young's, except that it employs an electron source in place of the light source and a suitable detector to record the pattern of electron intensities at a given distance from the slits (Fig 2.35).

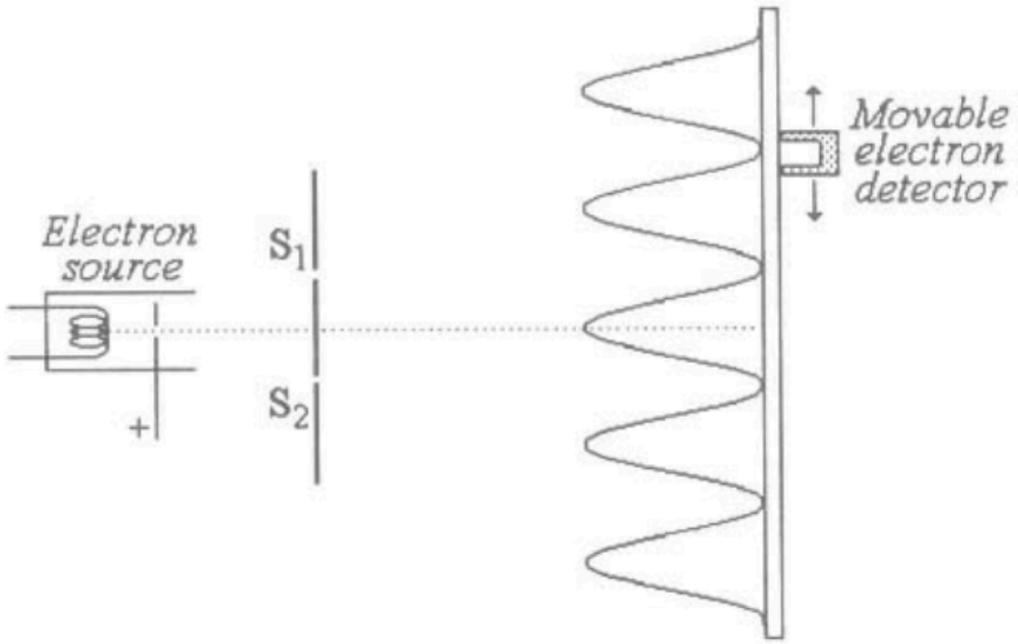


Fig 2.35 A double-slit experiment with electrons. When the two slits, S_1 and S_2, are kept open at the same time, an interference pattern of alternating maxima and minima is recorded by the movable electron detector. With just one slit open or if a detector is installed in the apparatus which determines through which slit each electron passes, a single broad diffraction band is recorded.

The two slits are alternative routes by which electrons can pass from the source to the movable electron detector.

If just one slit is open at any one time or if both are kept open all the time but a detector that records through which slit each electron passes is also installed in the apparatus, the route taken by each electron can be unequivocally ascertained.

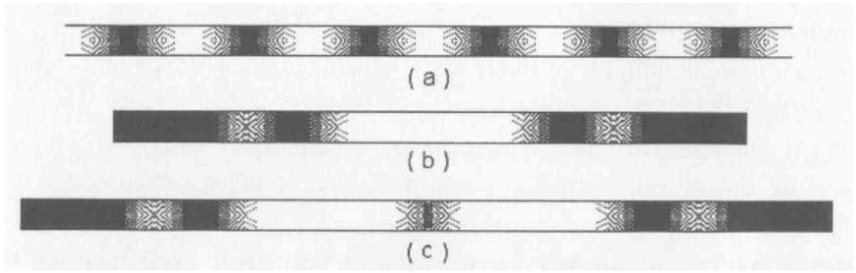
In this case, the probability, P_1, of an electron passing through the first slit is given by $|\Psi_1|^2$. and the probability, P_2, of it passing through the second slit by $|\Psi_2|^2$.

Overall, the probability, P_{12}, of an electron reaching the movable detector at a particular point is

$$P_{12} = P_1 + P_2 = |\Psi_1|^2 + |\Psi_2|^2 \tag{2.70}$$

and there is no interference.

The distribution of electrons recorded by the movable detector corresponds to two overlapping single-slit diffraction patterns (Fig 2.27c)



However, if both slits are kept open at the same time and no additional detectors are installed in the apparatus, there is no way of ascertaining through which slit each electron actually passes.

In this case, the wave function that describes the passage of an electron from the source to the movable detector is $\Psi = \Psi_1 + \Psi_2$, and the probability, P, that an electron reaches the movable detector at a particular point is

$$P = |\Psi|^2 = |\Psi_1 + \Psi_2|^2 = |\Psi_1|^2 + |\Psi_2|^2 + 2|\Psi_1||\Psi_2|\cos\phi \tag{2.71}$$

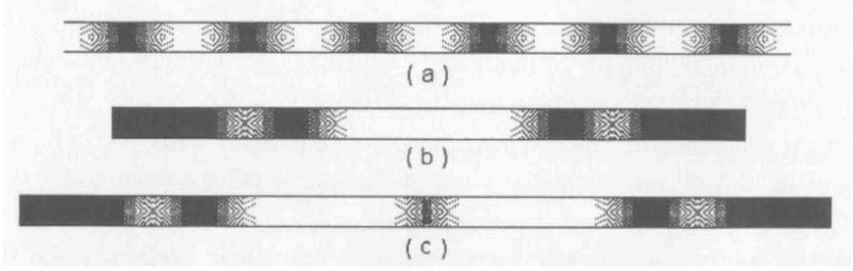
where ϕ is the phase difference between the wave functions at the point.

The term $2|\Psi_1||\Psi_2|\cos\phi$ is an interference term whose value varies with that of $\cos\phi$.

Many electrons reach those points where $\cos\phi = +1$ and which correspond to constructive interference of the wave functions.

In contrast, the chance of an electron reaching those points that correspond to destructive interference where $\cos\phi = -1$ is negligible.

Thus, a typical interference pattern of alternating intensity maxima and minima is recorded by the detector (Fig 2.27a).



In general, the alternate routes or stages by which a physical event can take place are delineated by the structure of the apparatus or system in which the event occurs.

Accordingly, every change in the system's dimensions, components or design will affect the event's overall wave function.

This accounts for the 'awareness' of the experimental set-up that we attributed to the photons in the description of the experiment in the previous chapter.

It is as though each of the particles involved in an event ascertains all the ways in which it can occur and thus becomes aware of the experimental set-up.

The overall wave function is formed from the appropriate addition of the wave functions of all the alternate routes available to the particles.

Despite the difficulty many people have in accepting these strange and non-intuitive ideas, the technique of ascertaining the outcome of events by means of wave functions has passed every single test and experimental assay to which it has been subjected, without exception.

It does work.

> *Throughout much of history, one of the major positions in the methodology of science has been what is sometimes referred to as the 'save the phenomenon' position...The basic tenet of this position is that judgements of the value of a scientific theory should be made in terms of only two criteria: the ability of the theory to save (account for or predict) the phenomena and the theory simplicity. Among the criteria viewed as inappropriate is the realist's claim that a theory must be true. For an advocate of the save the phenomena position, a theory is nothing more than a useful fiction or instrument to account for phenomena.*

Although quantum mechanics satisfies the first criterion, it does indeed 'save' the phenomena, its simplicity is not self-evident.

> *As man's investigation of the universe progressed to the almost infinitely large distances of interstellar space or to the almost infinitesimal magnitudes of atomic structures, it began to be realized that these other worlds could not be adequately described in terms of the bricks and mortar and plumbing of terrestrial architecture...Intensive research into the ultimate nature of our universe is thus gradually changing the meaning we attach to such words as 'explanation'or 'understanding'. Originally they signaled a representation of the strange in terms of the commonplace; nowadays, scientific explanation tends more to be a description of the relatively familiar in terms of the unfamiliar light in terms of photons; matter in terms of waves.*

2.4.5 The Wave Function of a Free Particle

A free particle is one that is not subject to any force; it is free to roam anywhere.

As such, there are no restrictions on its total energy; it can take any value.

The energy eigenvalues of a free particle -its energy levels- merge into an unbroken continuum.

The wave function of a free particle cannot be such that it assigns the same probability to all points in space-time.

The particle must actually be somewhere and its matter wave at that place must give a the greatest probability.

Hence, its wave function cannot be a uniform infinite sinusoidal wave for such a wave assigns the same probability to every point in space.

The fact that the wave function of a free particle cannot be a uniform infinite sinusoidal wave is also confirmed by considerations of its phase velocity.

The phase velocity, v, of a sinusoidal wave is equal to the product of its frequency and wavelength.

According to De Broglie's hypothesis, the frequency of a particle's matter wave is given by $f = E/h$, where $E = mc^2$ is the particle's total energy.

Likewise, its wavelength is given by $\lambda = h/mu$, where u is the particle's velocity.

Combining these three expressions gives

$$v = f\lambda = \frac{mc^2}{h} \cdot \frac{h}{mu} = \frac{c^2}{u} \tag{2.72}$$

Since, by reason of special relativity, the particle's velocity $u < c$, it follows from equation (2.72) that if the matter wave was uniform and sinusoidal, its phase velocity $v > c$, i.e., the wave would propagate at a speed greater than that of light.

This of itself presents no problem, for no physical quality is actually moving; matter waves as such can neither be measured nor interpreted in terms of any experiment.

But it does mean that a single uniform sinusoidal matter wave cannot represent a free particle, for its phase velocity, v, cannot be the velocity of a real object.

A free particle's matter wave is in fact a wave group or wave packet, that is built up from the superposition of an infinity of sinusoidal matter waves of different amplitudes and wavelengths (Fig 2.36).

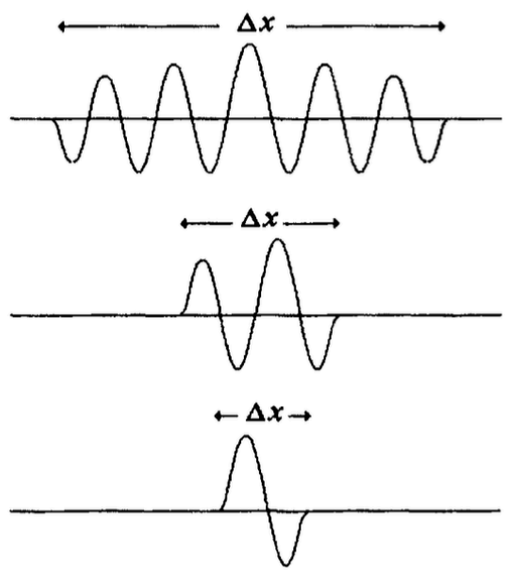


Fig 2.36 Th matter waves of free particles.The more confined the wave group the smaller the uncertainty, Δx , in the particle's position, but the greater will be the uncertainty, $\Delta\lambda$, in the wavelength, i.e., in the particle's linear momentum.

The wave group is the result of the superposition of the component waves.

The waves interfere destructively everywhere except in the region where the particle actually is.

The velocity of the particle is not that of the component sinusoidal matter waves but the group velocity, i.e., the velocity with which the group of waves propagates.

Thus, a free moving particle is analogous to a signal transmitted by a matter wave.

The inherent uncertainty in a particle's position depends on the width of its wave group, Δx.

The smaller the wave group the lower the uncertainty in the particle's position, and vice versa.

However, the smaller the wave group, the less well defined is its wavelength and the greater, therefore, its uncertainty, $\Delta\lambda$.

This, in turn, implies a greater uncertainty in the particle's linear momentum.

Hence, exactness in fixing the particle's position is achieved at the expense of the accuracy with which its momentum can be measured, which is specifically what Heisenberg's uncertainty principle, $\Delta x\ \Delta p_x \approx \hbar$,asserts.

The time a wave packet takes to move a distance equal to its width is

$$\Delta t = \Delta x / u \tag{2.73}$$

This is the time required for the particle to move a noticeable distance, i.e., for an appreciable change to occur in the system.

The uncertainty in the particle's energy, ΔE, is given by the energy-time uncertainty relationship, $\Delta E = \hbar / \Delta t$.

Hence, in general, the longer it takes for an appreciable change to occur in a system, the more sharply observable is its energy, and vice versa.

2.4.6 Quantum Mechanics - Schrodinger's Equation

Classical mechanics cannot be employed in systems in which the wave character of particles is manifested.

The attempt to assign trajectories to particles that show pronounced wave properties does not yield a correct description of their actual behavior.

A new mechanics is needed for such systems; this new mechanics is called Quantum Mechanics or Wave Mechanics.

In content, this new mechanics is identical to the matrix mechanics developed by Heisenberg.

The equations of motion of classical mechanics ensue from the three postulates Newton called the "Axioms or Laws of Motion".

In the one dimensional case, the second of these axioms-Newton's second law-can be expressed as the familiar differential equation

$$F = m\frac{d^2x}{dt^2} \tag{2.74}$$

This equation has no a priori foundation.

However, its solutions are found to be consistent with experiment in many physical situations. Its value and validity spring from this accord.

In place of the equations of motion of classical mechanics from which the particle's exact position in space at every moment can be computed, quantum mechanics provides wave functions that contain all that can be known about the particles subject to the uncertainty principle.

The wave functions of quantum mechanics can be derived from a fundamental differential equation known generally as Schrodinger's Equation, that has the same status in quantum mechanics as that held by Newton's Laws in classical mechanics.

It too is a postulate that has no *a priori* standing and is only as good as the consistency between its solutions and experiment.

Whatever other properties it might have, we would expect the Schrodinger equation to incorporate the following fundamental principles:

(i) **The conservation of energy: this principle is so basic that its exclusion is unthinkable.**

(ii) **De Broglie's hypothesis: quantum mechanics is specifically intended for particles that show distinctive wave properties.**

(iii) **The principle of linear superposition: the interference of matter waves must be inherent in the equation.**

We shall now show how a differential equation that embraces these fundamental principles can be constructed.

The conservation of a particle's energy is defined by the equation

$$K = E - U \tag{2.75}$$

where K is the particle's kinetic energy, E its total mechanical energy and U its potential energy.

In terms of its mass, m, and its linear momentum, p, the kinetic energy of a particle of is given by:

$$K = p^2/2m \tag{2.76}$$

Substituting the De Broglie equation, $p = h/\lambda$, in this relationship gives

$$K = h^2/2\lambda^2 m = E - U \tag{2.77}$$

Let us assume, for simplicity's sake, that the time independent part of the particle's one dimensional wave function, ψ, has the form

$$\psi = A \sin kx \tag{2.78}$$

where $k = 2\pi/\lambda$.

The second derivative of this function is:

$$\begin{aligned}\frac{d^2\psi}{dx^2} &= -k^2 \sin kx \\ &= -k^2\psi \\ &= -\frac{4\pi^2}{\lambda^2}\psi\end{aligned} \tag{2.79}$$

Eliminating λ^2 from equations (2.77) and (2.79) gives:

$$\frac{d^2\psi}{dx^2} = -\frac{8\pi^2 m}{h^2}(E-U)\psi \tag{2.80}$$

This equation is the one dimensional form of the general time independent Schrodinger equation.

It is not the only equation we could have constructed; others could have been put together that incorporate the same three principles.

However, it is the solutions of this equation and of its three-dimensional analogues that are found to be consistent with experimental results.

In principle, the procedure for solving a system's Schrodinger equation is quite straightforward.

The mathematical function that describes the system's potential energy in terms of its space and time coordinates is substituted in the equation in place of U and the resulting differential equation is then solved subject to the normalization requirement that the particle must be somewhere.

However, though the procedure sounds simple, in most cases, the mathematics required for the solution of this differential equation are anything but straightforward.

2.4.7 Quantum Mechanics Potential Wells

The simplest example of a quantum system is that of a particle for which U = 0? In the one-dimensional time independent case this gives for the particle's Schrodinger equation:

$$\frac{d^2\psi}{dx^2} = -\frac{8\pi^2 m}{h^2} E\psi \tag{2.81}$$

We will illustrate the solution of this equation by considering the case of a particle trapped in a one dimensional infinite potential well of length L, a box with infinitely impenetrable sides (Fig 2.39).

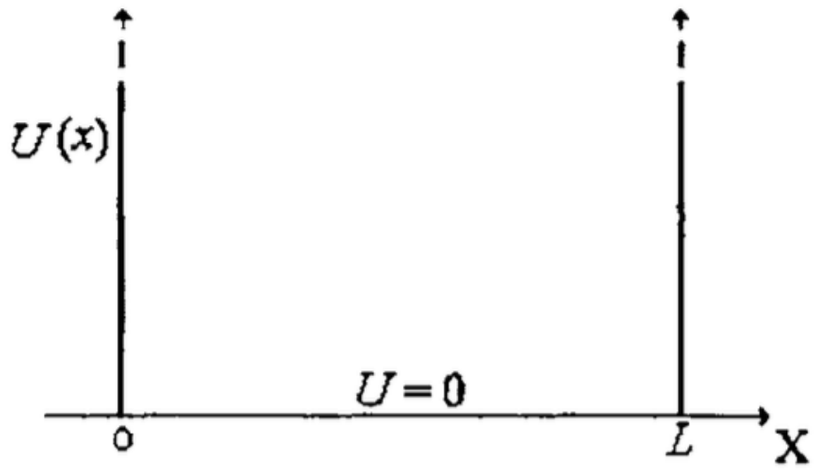


Fig 2.37 A one-dimensional infinite potential well of width L. A particle situated inside the well can never escape, the height of the walls being infinite. The particle can only fly to and fro in the space between the walls.

We could solve equation (2.81) by the standard mathematical technique.

But instead, let us use our physical intuition.

Inside the well, the potential energy of the particle is everywhere zero; at the walls of the well it rises to infinity.

Lacking the energy required to mount the infinite energy barrier that the walls of the well represent, the particle can never escape. It is trapped in the well.

This being so, its matter wave must reduce to zero at the walls of the well.

This boundary condition must be reflected in the solution of the particle's Schrodinger equation.

It suggests that the trapped particle's matter wave has the form of a standing wave, of the type encountered in a vibrating wire stretched between two fixed points in which there are nodes at each end.

The wavelengths, λ_n, that can produce such standing waves in a stretched wire of length L are given by:

$$\lambda_n = {}^{2L}\!/_{n} \quad (n = 1, 2, 3, \ldots) \tag{2.82}$$

i.e., the allowed wavelengths are 2L/1, 2L/2, 2L/3, ……

The integer n, which characterizes the pattern of the trapped particle's matter wave, is called a *quantum number*.

Each value of n corresponds to a different vibrational mode of the matter wave (Fig 2.40).

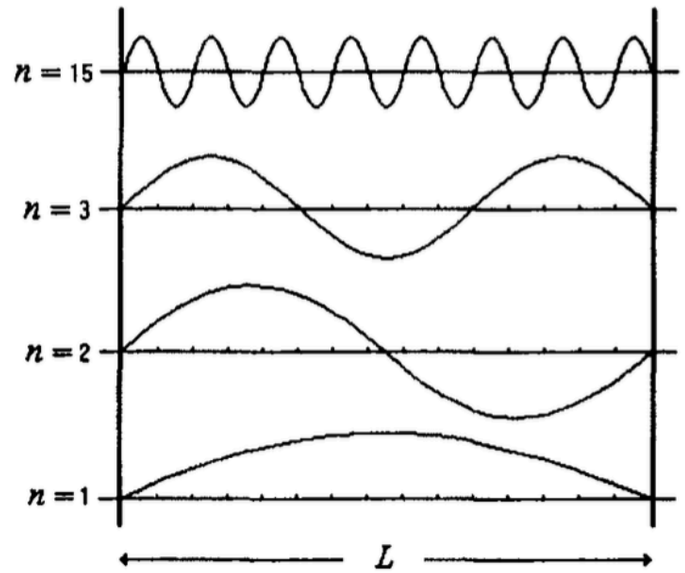


Fig 2.38 The standing waves produced in a stretched wire of length L for different values of the integer n. The wave functions of a particle trapped in an infinite potential will have a similar form.

Since the particle's potential energy is zero, its total energy is its kinetic energy: E = K.

Substituting the trapped particle's allowed wavelengths, $\lambda_n = 2L/n$, in the quantum mechanical expression for its kinetic energy, $K = h^2/2\lambda^2 m$, gives the following expression for its total energy, E_n:

$$E_n = \frac{n^2 h^2}{8mL^2} \quad (n = 1, 2, 3, \ldots) \tag{2.83}$$

These are the energy eigenvalues that define the energy levels of the particle trapped in the one dimensional infinite potential well (Fig 2.39).

Delineation of the energy of a particle in a three-dimensional well would require three quantum numbers.

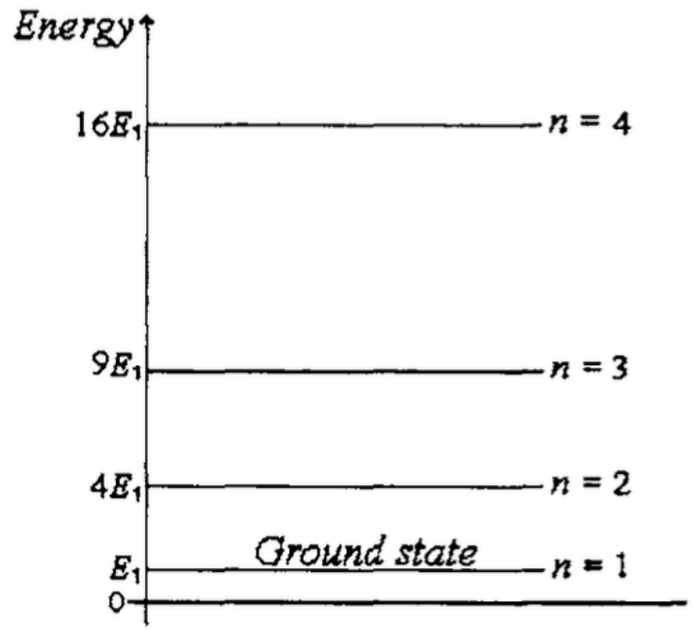


Fig 2.39 The energy levels of a particle trapped in a one dimensional infinite potential well. Even in the ground state., n = 1, the particle's energy is greater than

In complete contrast to the situation in classical or relativistic mechanics, where the energy of a trapped or bounded particle can change continuously, in quantum mechanics the energy of a confined particle is quantized, i.e., the particle's energy can only change by certain discrete amounts.

Thus, the energy, E_n, of a particle trapped in a one dimensional infinite potential well always equals an integer multiple of the basic quantum,

$$E_1 = \frac{h^2}{8mL^2}$$

Furthermore, the particle's energy can never be less than a single basic quantum; its quantum number, n, can never be less than unity.

Thus, the particle can never be completely at rest even at the absolute zero of temperature.

In the case of a small macroscopic particle such as a grain of sand of mass ~10^{-6}kg trapped in a well of length lm, the size of its basic quantum of energy is approximately 10^{-62}J.

This is such a minute quantity of energy, that it is in practice undetectable, and so changes in the particle's energy are to all intents and purposes continuous and can be properly described by classical mechanics.

This is a further example of the correspondence principle.

On the other hand, the basic quantum of energy of a particle the size of an electron trapped in a potential well of length 10^{-10}m (the approximate size of an atom) is approximately 30eV.

This is a significant quantity of energy; it is the energy acquired by an electron accelerated through a potential of 30V.

Consequently, the quantization of the energy of an atom's electrons is a central feature in the structure of the atom and it is in such systems that quantum mechanics finds most of its applications.

The eigenvalues of the trapped particle's linear momentum are obtained by substituting the expression for its energy eigenvalues

$$E_n = \frac{n^2 h^2}{8mL^2}$$

in the formula

$$p = \pm\sqrt{2mE_n}$$

$$p_n = \pm \frac{n\pi\hbar}{L} \quad (n = 1, 2, 3, \ldots) \tag{2.84}$$

The particle is moving to and fro and so its average linear momentum is zero.

The complete solution to the particle's Schrodinger equation is obtained by its integration subject to the following two conditions:

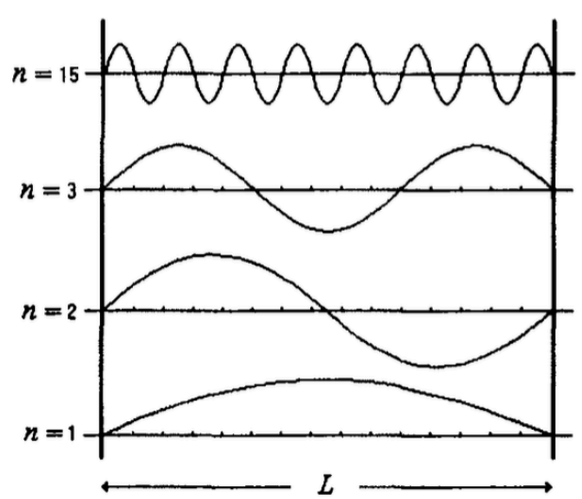
(i) the boundary condition that the wave function must equal zero at the well's walls, i.e., $\psi(x) = 0$ when $x = 0$ or L ;

(ii) the normalisation condition, $\int_0^L \left|\psi^2(x)\right| dx = 1$.

This gives for the particle's wave function

$$\psi(x)_n = \sqrt{2/L} \cdot \sin\left(n\pi x/L\right) \quad (n = 1, 2, 3, \ldots) \tag{2.85}$$

The standing waves illustrated in Fig 2.38 correspond to this equation.



The corresponding probability densities, $|\psi(x)|^2$, are shown in Fig 2.40.

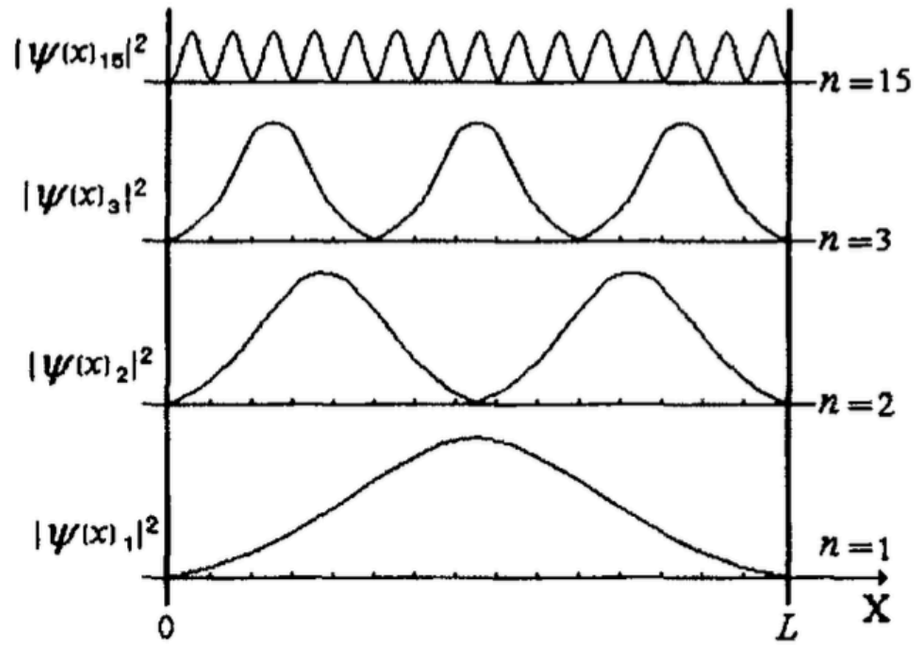


Fig 2.40 The probability densities for various values of the quantum number n. The probability of finding the particle close to a point is proportional to the probability density at the point. Note that in every case the probability of finding the particle at the walls of the infinite potential well is zero

In general, the total energy of a particle that is confined to a specific region of space, such as the particles constituting an atom, can only take certain discrete values.

The particles are restricted to certain stationary states in which their energy is *quantized*.

The energy eigenvalues of these stationary states, i.e., their energy levels, are sharply observable.

The *eigenfunctions* that describe the stationary states the particle can occupy have the form

$$\Psi_S(x,y,z,t) = \psi(x,y,z)e^{-\frac{iEt}{\hbar}} \tag{2.86}$$

where E is the energy eigenvalue of the stationary state, i.e., its energy level.

The probability of finding the particle in a particular stationary state does not change with time.

Hence, the probability density of a stationary state

$$\left|\psi(x,y,z)e^{-\frac{iEt}{\hbar}}\right|^2$$

is constant.

This is the meaning quantum mechanics gives to the concept of a stationary state.

2.4.8 The Tunnel Effect

Infinite potential wells do not exist in reality; real energy barriers are of a finite height (Fig2.41).

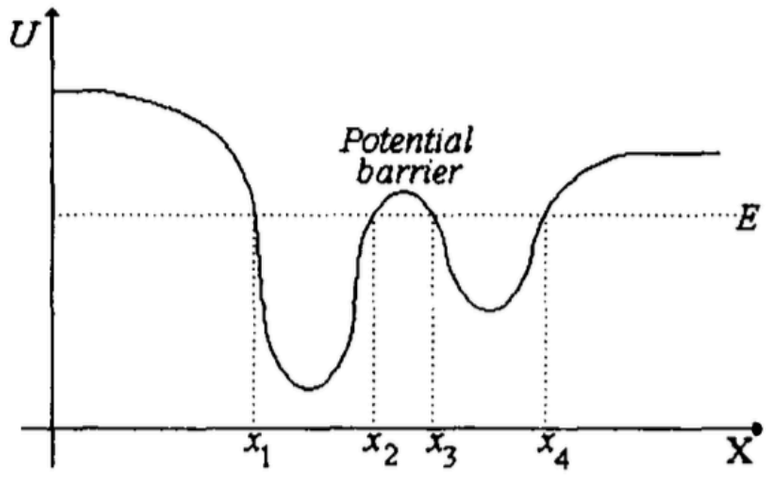


Fig 2.41 A hypothetical potential energy diagram. Classically, a particle with energy E can oscillate in either of the two potential wells, i.e., between x_1 and x_2 or between x_3 and x_4. However, it has insufficient energy to pass over the potential barrier between x_2 and x_3; it is trapped in one or other of the wells. Quantum mechanics asserts that there is a non-zero probability that the particle will 'tunnel' through the barrier.

In classical physics, a particle can only escape from a finite potential well if it acquires sufficient energy to climb over its walls; there is no way it can get out with less energy.

A ball rolling down one side of a bowl can rise to the same height on the opposite side, but no higher.

However, this is not the case in quantum mechanics.

Because of the inherent uncertainties in its position, the wave function of a particle does not equal zero at the walls of the well, and so there is a non-zero probability of finding the particle there.

Consequently, the possibility exists, slight though it may be, that the particle will penetrate or 'tunnel' through the walls, i.e., through the energy (potential) barrier keeping it in the well.

Thus, it can escape from the potential well even when it has insufficient energy to climb over its walls. This phenomenon is called the *tunnel effect* and is a characteristic of quantum behavior

Most of the particles striking the wall of a potential well will be reflected from it.

Those that do get across to the other side are never observed inside the wall, where they would have a negative kinetic energy: $K_{wall} = E - U < 0$ (Fig 2.42).

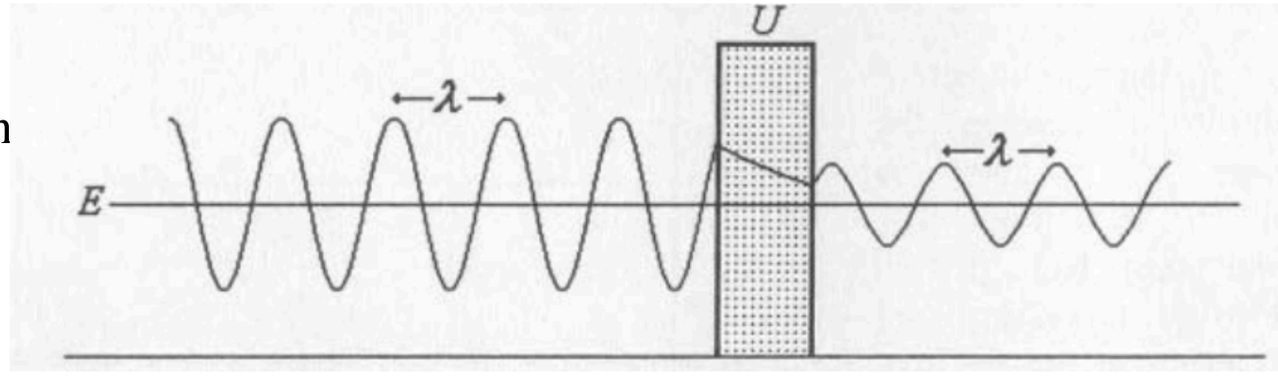


Fig 2.42 The tunnel effect. The wave function of a particle which strikes the barrier from the left. Its energy, E, is less than the height of the barrier, U. Nevertheless, the wave function to the right of the barrier is not zero, though its amplitude is less than on the left. As a result, there is a non-zero probability of finding the particle to the right of the barrier.

Many phenomena are explained by the tunnel effect.

For example, when two copper wires are in contact with one another, an electric current flows between them even though each is covered by a thin layer of nonconducting copper oxide.

The conducting electrons simple ‘tunnel’ through the oxide barrier.

The emission of a particles (Helium nuclei) from the nuclei of radioactive atoms is another example of quantum tunneling.

Even though the α particles have insufficient energy to overcome the barrier that holds nucleons inside the atom’s nucleus, there is a non-zero probability that they will appear outside the nucleus.

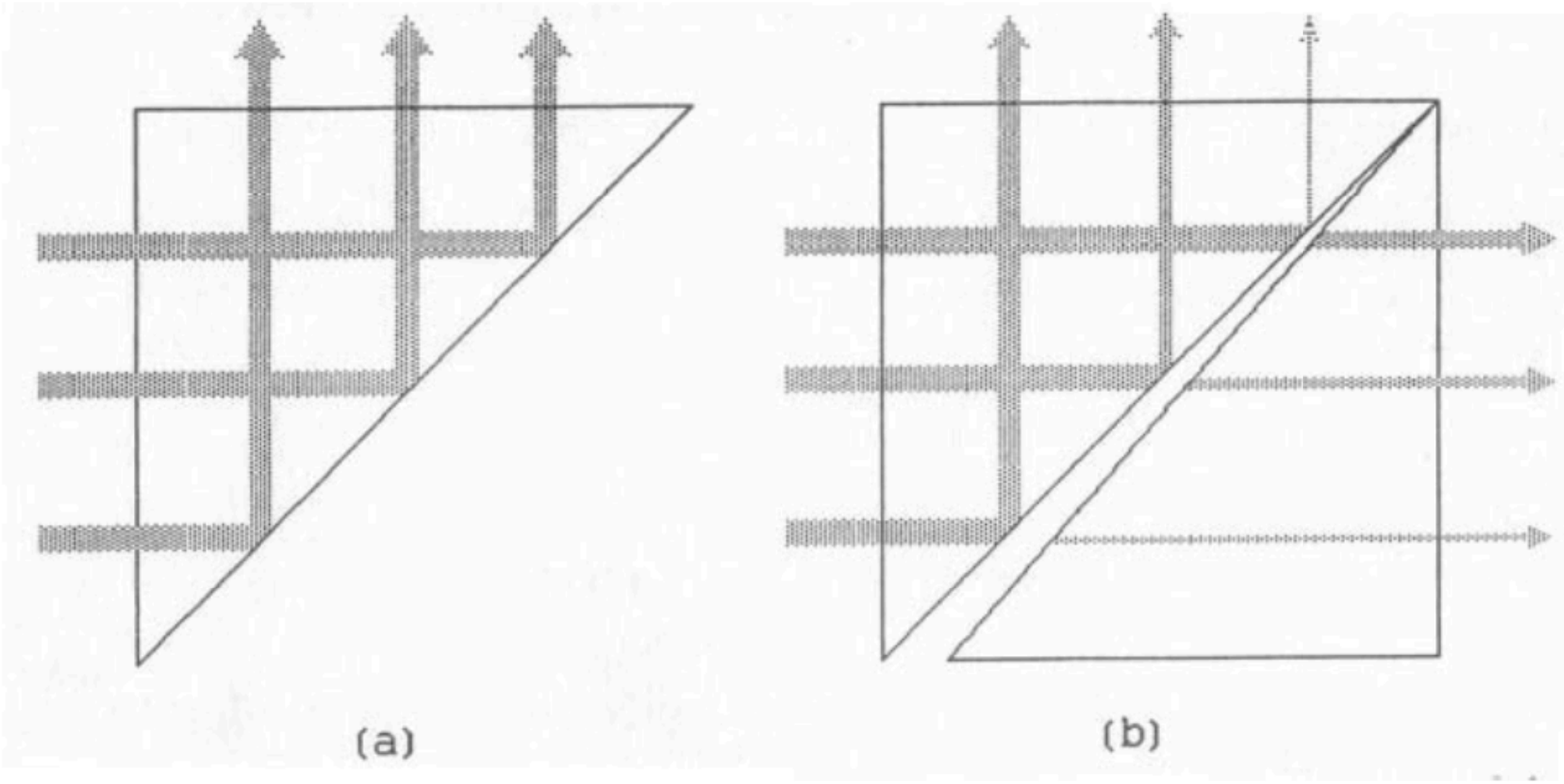
The phenomenon of barrier penetration also occurs in macroscopic systems traditionally described in terms of waves.

For example, how do the waves inside a glass prism ‘know’ that there is air on the outside and that they should be totally reflected internally (Fig 2.43a)?

It must be that on reaching the edge of the prism, the waves continue some small distance into the air in order to ‘feel’ what is outside.

If a second prism is placed close to, but not actually touching the first, light reaches and enters the second prism (Fig 2.43b).

Fig 2.43 (a) Total internal reflection. (b) The spacing between the prisms is of the order of a few wavelengths. Light 'jumps' across the gap and appears in the second prism. No light is detected in the gap.



However. no light is detected in the space between them.

The phenomenon can also be demonstrated with water and sound waves.

Now we finish Part 2 with some detailed derivations of quantum properties.

2.5.1 The Kinetic Energy and Linear Momentum of a Particle

Eliminating the particle's velocity, u, by simple algebra, from the familiar classical equations $K = {mu^2}/{2}$ and $p = mu$, for the kinetic energy and linear momentum of a particle of mass m respectively, gives

$$p = \sqrt{2mK}$$

2.5.2 The Wave Function of a Trapped Particle

The general time-independent Schrödinger equation for a particle of mass m and potential energy $U(x,y,z)$ is

$$\frac{-h^2}{8\pi^2 m}\left(\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}\right)=(E-U)\psi$$

The particle's wave function, ψ, and energy levels, E, are obtained from the solution of this equation.

In the one-dimensional case, equation (2.104) reduces to:

$$\frac{d^2\psi}{dx^2}=\frac{-8\pi^2 m}{h^2}(E-U)\psi$$

The potential energy of a particle trapped in an infinite potential well is zero: $U = 0$.
At the walls of the well the system's potential energy rises to infinity: $U = \infty$.
Substituting $U = 0$ above equations gives:

$$\frac{d^2\psi}{dx^2} = \frac{-8\pi^2 mE}{h^2}\psi$$

Making the substitution

$$k^2 = \frac{8\pi^2 mE}{h^2}$$

reduces equation (2.106) to the familiar harmonic form:

$$\frac{d^2\psi}{dx^2} = -k^2\psi$$

The general solution of this equation is:

$$\psi(x) = A\sin kx + B\cos kx$$

The particle is trapped inside a potential well of width L, and so its wave function must equal zero at the well's walls. This requirement defines the following boundary conditions for the wave function: that $\psi(x) = 0$ when (i) $x = 0$; (ii) $x = L$. The first condition gives:

$$\psi(0) = A\sin(0) + B\cos(0) = 0$$

Satisfying this equation requires that $B\cos(0) = 0$, i.e., that $B = 0$. This reduces this equation to the form:

$$\psi(x) = A\sin kx$$

The second boundary condition gives

$$\psi(L) = A\sin kL = 0$$

Satisfying this equation requires that $\sin kL = 0$, i.e., that

$$kL = n\pi$$

or

$$k^2 = \frac{n^2\pi^2}{L^2}$$

where $n = 1, 2, 3, \ldots$. Substituting for above equation gives for the energy of the trapped particle:

$$E = \frac{n^2h^2}{8mL^2}$$

In order to complete the solution of the wave equation, we adopt Born's interpretation of the wave function, i.e., that the probability P of finding the particle in the region between x and $x + \Delta x$ is given by the square of the absolute value of the wave function:

$$P = |\psi(x)|^2 \cdot \Delta x$$

The particle must be somewhere inside the potential well, i.e., the probability of finding it somewhere in there must equal unity. This defines the normalization requirement

$$\int_0^L |\psi(x)|^2 \, dx = 1$$

Substituting $\psi(x) = A\sin kx$ from above equation gives

$$A^2\int_0^L \sin^2 kx \cdot dx = A^2\int_0^L \frac{1-\cos 2kx}{2}\,dx$$

$$= A^2\left(\frac{L}{2} - \frac{\sin 2kL}{4k}\right)$$

$$= 1$$

However, according the equations , $\sin 2kL = 0$. Hence,

$$A = \sqrt{2/L}$$

Finally, using above equation and substituting for A in above equation gives:

$$\psi_n(x) = \sqrt{2/L}\sin\left(n\pi x/L\right)$$

The waves illustrated in Fig 2.38 match this equation.

Part 3 The Nuclear Atom

The word atom is of Greek origin - ατομοσ - meaning indivisible.

The ancient Greek philosophers Leucippus, Epicurus and Democritus supposed that the properties of materials were determined by the different shapes and forms of their atoms.

According to this notion, iron atoms are solid and strong with hooks that lock them into a solid; water atoms are smooth and slippery; salt atoms, because of their taste, sharp and pointed, and air atoms light and whirling, pervading all other materials

Other ancient Greek philosophers, such as Empedocles, Heraclitus and Aristotle, considered all matter to be composed of the four elements, earth, air, fire and water, in varying proportions.

According to this notion, lead differs from gold only in the proportions of the four elements it contains and so the problem of converting lead into gold is just a matter of correctly blending the right combination of elements.

The search for this recipe, which, with the addition of a pinch of a very special and rare spice called the philosopher k stone, would turn base metals into gold and enrich its finder beyond his wildest dreams, occupied the minds and time of many learned men for more than a thousand years.

Atoms were forgotten in the frenzy of the alchemists' search for the formula and the essence.

Modern ideas about the structure of matter originated in the 17th century.

In *The Sceptical Chymist*, published in 1661, Robert Boyle laid down the modern criterion of an element: a substance that cannot be decomposed chemically.

This definition set the agenda for much of experimental chemistry for the next 200 years.

A counterpart to this was the renascence of the atom brought about by the writings of Descartes, Gassendi and Newton.

Anticipating kinetic theory, Hooke even suggested that the source of the 'spring of the air' was the collisions of its atoms with the walls of the container

Summing up the general belief in atoms, Newton wrote in the final query he appended to his book *Opticks:*

> ***All things being considered, it seems probable to me that God in the beginning formed matter in solid, massy, hard, impenetrable, moveable particles, of such sizes and figures, and with such other properties, and in such proportion, as most conduced to the end for which He formed them; and that these primitive particles, being solids, are incomparably harder than any porous bodies compounded of them; even so very hard, as never to wear or break into pieces.***

In his Essay on Human Understanding, the philosopher Locke re-examined the problem of how atoms could account for the many and varied properties of material objects.

He distinguished between primary qualities such as motion, size and shape, that he argued are inherent in the material's atoms, and secondary properties such as color, taste and smell, that exist only in the mind of the observer and result from an interaction between the atoms in the object and those in the sense organs of the observer.

It was this dichotomized interpretation of perception, i.e., that though an object's shape is actually seen its color is only in the mind, that marked the severance of science from philosophy.

Whereas philosophy continued to explore the mind and its ideas, experimental science set out to examine the physical world and the atoms from which it is composed.

Boyle's empirical law, that the volume of a gas is inversely proportional to its pressure, was given a sound mathematical basis by Daniel Bernoulli in 1738, when he showed that it could be derived from mechanical considerations of the collisions between the particles of the gas and the walls of its container.

In so doing, Bernoulli founded the classical kinetic theory of matter.

Assuming material substances and physical systems to be composed of minute particles, this theory seeks to explain their macroscopic properties in terms of the Newtonian motion of their constituent particles.

Direct evidence of the motion of the these particles was first obtained by Brown in 1827, when he observed the random movement of fine grains of pollen in an aqueous suspension using the recently invented achromatic microscope.

In 1803 Dalton published the first comprehensive atomic theory, *In A New System of Chemical Philosophy* he proposed that the elements are comprised of indestructible atoms, all the atoms of a particular element being identical.

The notion of the molecule, compounded of atoms, as the basic and characteristic chemical unit of matter was proposed by Avogadro in 1811 and, in 1812, Berzelius first suggested that the forces binding the atoms together were electrical in origin.

Modern atomic theory began with the discovery of the electron.

This discovery opened the way for the integration of the mechanical, the chemical and the electrical properties of the atom into a single atomic theory-the nuclear model of the atom

The notion of an essentially indestructible and indivisible atom remains fundamental to much of modem chemistry and biology, though, as we will see in Part 5 of this discussion, by the middle of the twentieth century the techniques of splitting atoms had been mastered by modem physics with devastating results.

Part 3.1 The Structure of the Atom

3.1.1 The Thomson Model of the Atom

Experiments carried out following their discovery in 1897 showed that electrons could be obtained, by one means or another, from just about any material.

For example, by the action of ultra-violet light on metals, by heating metal wires, by the ionizing action of X-rays or from the spontaneous disintegration of radio-active substances.

Evidently, electrons are a component of all neutral atoms, the counterbalancing positive charge being carried by one or more of their other constituent parts.

However, since the electron's mass is just 0.05% that of the hydrogen atom, itself the smallest of all atoms, the contribution the electrons make to the overall mass of an atom is minuscule.

Most of an atom's mass must be associated with its positively charged constituent.

In the light of this, J.J.Thomson proposed in 1904 that a neutral atom is a sphere of homogeneous positively charged material, in which negatively charged electrons are scattered like "the plums in a pudding".

The positive charge attracts the electrons to the centre of the sphere, however, in atoms containing more than a single electron -multi-electron atoms- this attraction is balanced by the mutual repulsion between the electrons.

At equilibrium, a balanced distribution of the electrons results (Fig 3.1).

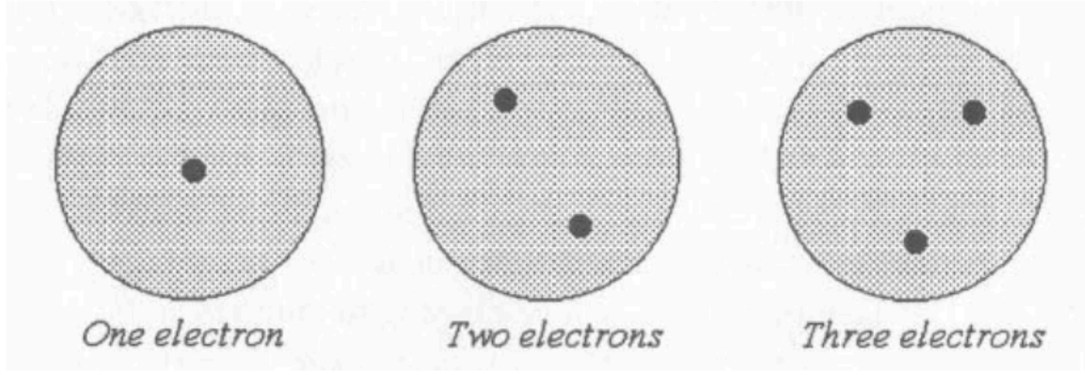


Fig 3.1 The Thomson model of the atom; the plum-pudding atom. In multi-electron atoms, the attractive force on the electrons is balanced by their mutual repulsion and a stable configuration results.

Calculations showed that the electrons in multi-electron atoms would arrange themselves in stable concentric rings and that, on certain assumptions, even in rings of eight, reminiscent of the groups of eight in Mendeledeev's periodic table of the elements.

As the number of electrons in an atom increased, more rings would be formed; the electrons in large multi-electron atoms would be arranged in a series of concentric rings.

Thomson went on to suggest, that the number of electrons in the outermost ring would determine the atom's electrovalency and was thus related to the element's position in the Periodic Table.

On this basis, the properties of the series of elements

He	Li	Be	B	C	N	O	F	Ne
Ne	Na	Mg	Al	Si	P	S	Cl	A

could be explained as follows:

> ***The first and last element in each of these series has no valency [their outermost ring has an unalterable number of electrons], the second is a monovalent electropositive element, the last but one is a monovalent electronegative element, the third is a divalent electropositive element, the last but two a divalent electronegative element, and so on.***

However, Thomson met with great difficulties when he tried to explain the emission spectrum of atomic hydrogen in terms of his model.

It can be shown, that if the electron in a Thomson hydrogen atom is disturbed from its equilibrium position at the center of the atom as the result of a collision or some other excitation, it will oscillate harmonically at a single frequency.

The excited electron constitutes an electric oscillator, and in accordance with electromagnetic theory it should emit electromagnetic waves of a frequency equal to that of its oscillation.

Thus, the emission spectrum of hydrogen atoms should comprise just a single line (in the far ultra-violet).

However, the emission spectrum of atomic hydrogen actually comprises a large number of lines Fig 3.2. The Thomson model had no explanation for this.

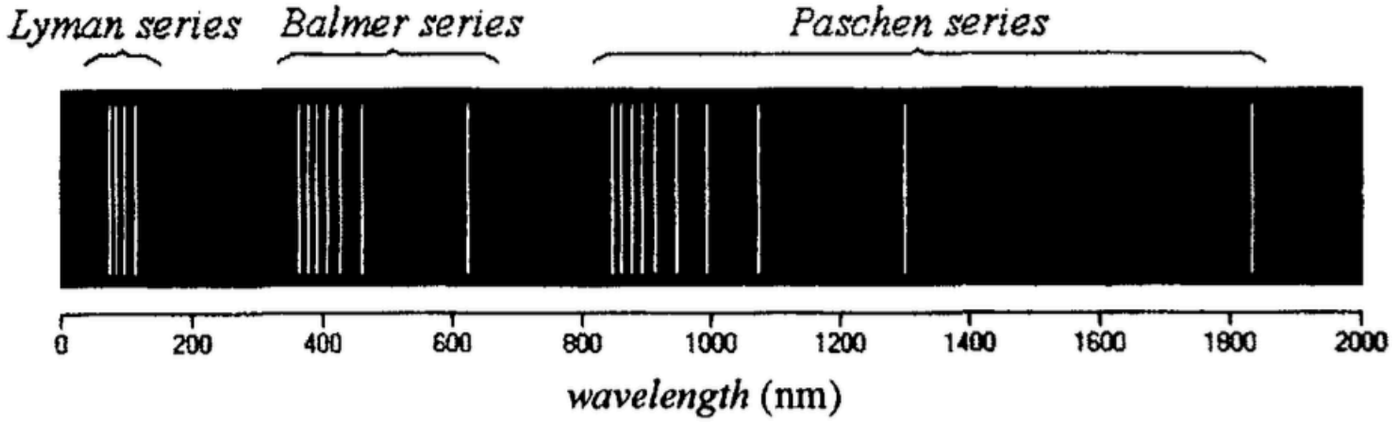


Fig 3.2 The emission spectrum of atomic hydrogen. The spectral lines appear indistinct series, named after their discoverers. Each series starts from a characteristic minimum wavelength, all the other lines in the series corresponding to longer wavelength.

To summarize. Thomson's model proposed that:

1. **Atoms comprise a sphere of positive material of uniform and 'normal' density $\left(\sim 10^3\,\mathrm{kg/m^3}\right)$ that accounts for virtually all of their mass and determines their spatial extension, e.g. their diameter.**
2. **At equilibrium, an atom's electrons are at rest and are distributed throughout the sphere of its positive component.**

These aspects of the model were investigated by Rutherford and his co-workers in 1910.

3.1.2 The Nuclear Atom

The discovery of X-rays and of the penetrating α and β radiations emitted by radioactive substances, provided investigators at the beginning of the twentieth century with powerful new tools for atomic research.

In the first experiments performed using these probes, a narrow collimated beam of β rays was aimed at a metal foil (β rays are high energy electrons, -1MeV).

The beam penetrated the foil and was found to be just slightly broadened when it emerged from the other side.

This broadening of the particle beam was attributed to the repulsion of the β particles by the electrons in the metal's atoms.

The number of electrons in each atom was estimated from the extent of the broadening.

Their number was found to depend on the nature of the atom; it varied from about 12 in the case of aluminum to about 100 in the case of gold.

These experiments confirmed that the number of electrons in an atom was not great enough for them to make a significant contribution to the atom's total mass.

In 1910, Geiger and Marsden investigated the scattering of a rays by metal foils.

A narrow collimated beam of a particles was aimed at a gold foil of approximately l μm thickness, i.e., about 10,000 atoms thick.

The particles in α rays are high energy helium nuclei, +6 MeV.

They carry a positive charge of +2e and have a mass that is about 7300 times greater than that of the electron.

Because of their relatively much greater mass, α particles are not significantly deflected from their paths by the electrons in the metal's atoms.

The α particles emerging from the gold foil were detected by the faint light flashes each produced on striking a glass plate coated with zinc sulphide; the plate was viewed through a low power microscope.

This particle detector or scintillation counter: as it was later termed, was mounted on an arm that could swivel in a circle around the metal foil.

The apparatus was enclosed in a vacuum chamber in order to avoid collisions between the a particles and air molecules (Fig 3.3).

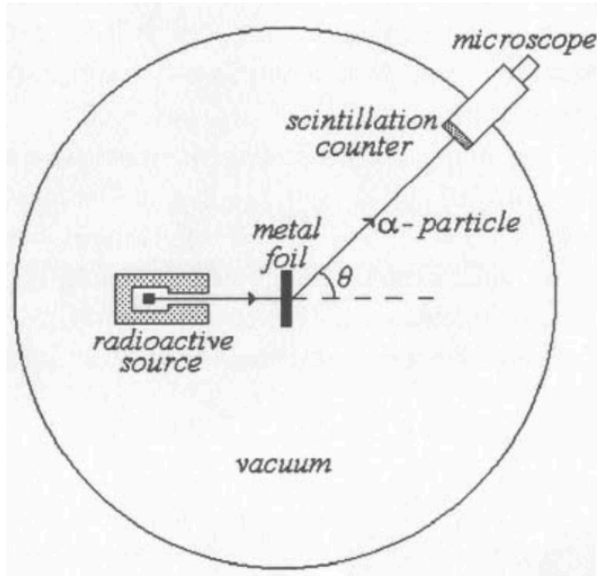


Fig 3.3 Geiger and Marsden's apparatus for investigating the scattering of α particles by a metal foil. The particles emerging from the foil at a particular angle θ were detected by the faint flashes of Iight produced when they struck the scintillation counter.

The number of a particles scattered through a particular angle, θ, during a given period of time, was found by counting the number of scintillations observed through the microscope at that angle.

In the first series of experiments, Geiger found that the average deflection of the a particles was 0.87°.

This result was in line with those that had been obtained from the scattering of β rays.

At Rutherford's suggestion, the experiment was repeated, perhaps with greater attention to details.

This time, the experimenters noticed that a few of the a particles, about 1 in 8000, were deflected through angles greater than 90°, i.e., they recoiled backwards.

When he was informed of this result Rutherford is reported to have said: "It is almost as incredible as if you had fired a fifteen inch shell at a sheet of tissue paper and it came back and hit you."

Such large deflections in the paths of the a particles could only result from a collision with a massive target or from an interaction with an electric or magnetic field of great strength.

That atoms are not impenetrable massive objects had been shown by the relative ease with which β rays and most of the α rays traversed the metal foils; this excluded the first possibility.

Thus, the deflections had to be electrical or perhaps possibly magnetic in origin.

The presence, inside the gold foil, of magnetic fields of the required intensity and orientation stretched the imagination, and so Rutherford concluded that the interaction had to be electrical (Fig 3.4).

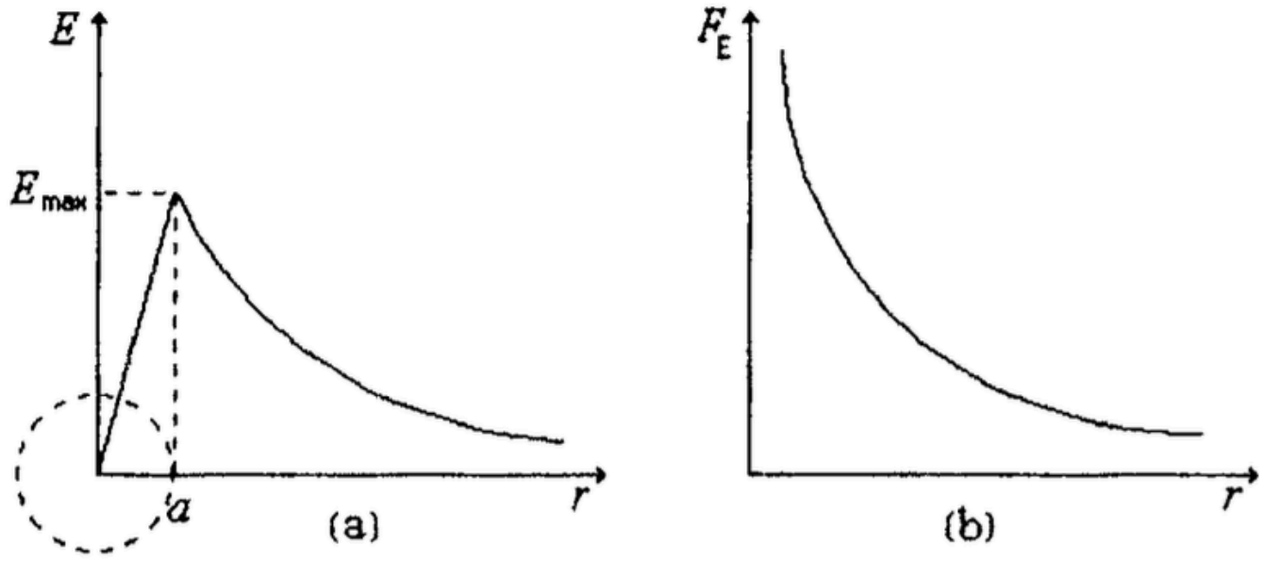


Fig 3.4 (a) The electric field, E, associated with a homogeneous charged sphere of radius a (a Thomson atom); the maximum field strength is at the sphere's surface. (b) The repulsive force, FE, between an a particle and the positive nucleus in a Rutherford atom as a function of the distance, r, between them; the closer the encounter between the particles, the greater the repulsive force between them becomes.

Confronted with these results, Rutherford abandoned the Thomson model. He pointed out that:

> *the theory of J.J.Thomson does not admit of a very large deflection in traversing a single atom, unless it is supposed that the diameter of the sphere of positive electrification is minute in comparison with the diameter of the sphere of influence of the atom.*

After considering the dynamics of the problem, Rutherford suggested that the positive component of the atom is localized in a very small volume, which he called the atomic nucleus, whose diameter is of the order 10^{-14}m.

Only by adopting this hypothesis could the very large deflections experienced by the a particles be explained (Fig 3.5).

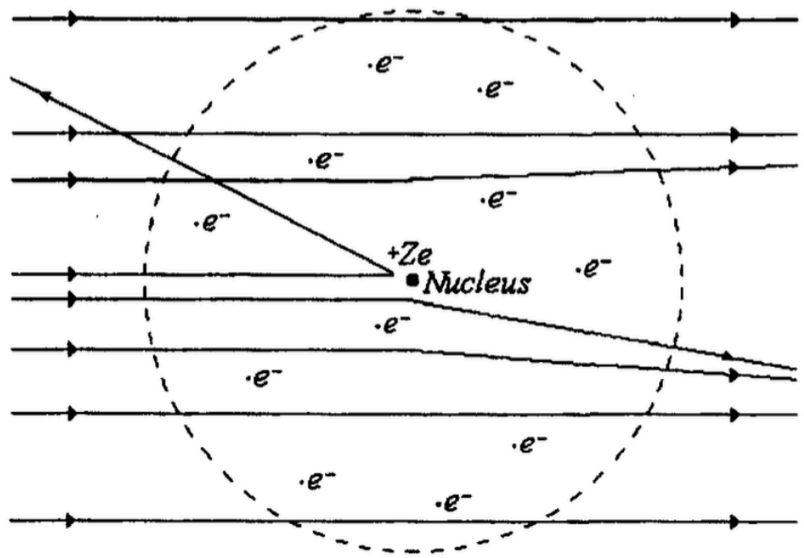


Fig 3.5 The scattering of a beam of a particles by a nuclear atom (not to scale). The atom's size, i.e., the diameter of its sphere of influence, is determined by its electrons. Most of the a particles pass through the atom without any noticeable deflection. Only those passing close to the nucleus, which carries a charge of +Ze, where Z is the number of electrons in a neutral atom, are significantly deflected. A particle moving very close to the nucleus may even be deflected backwards.

In Rutherford's atom, the diameter of its sphere of influence is determined by its electrons.

These gird the nucleus at a relatively great distance from it: about $l0^{-10}$m.

Thus, the diameter of the nuclear atom is some five orders of magnitude larger than that of its nucleus; the nuclear atom is largely empty space.

Since virtually all the atom's mass is concentrated within its minute nucleus, the density of the nuclear material must be immense.

Calculations showed it to be about 10^{14} tons per cubic meter; this is a hundred billion times greater than the density of water.

The electric field close to the surface of the nucleus is about 10^{21}V/m , i.e., some 10^{8} times greater than that at the surface of the Thomson atom (ignoring the effect of the electrons).

Rutherford explained that the scattering of the a particles through large angles resulted from a close encounter between an a particle and an atomic nucleus.

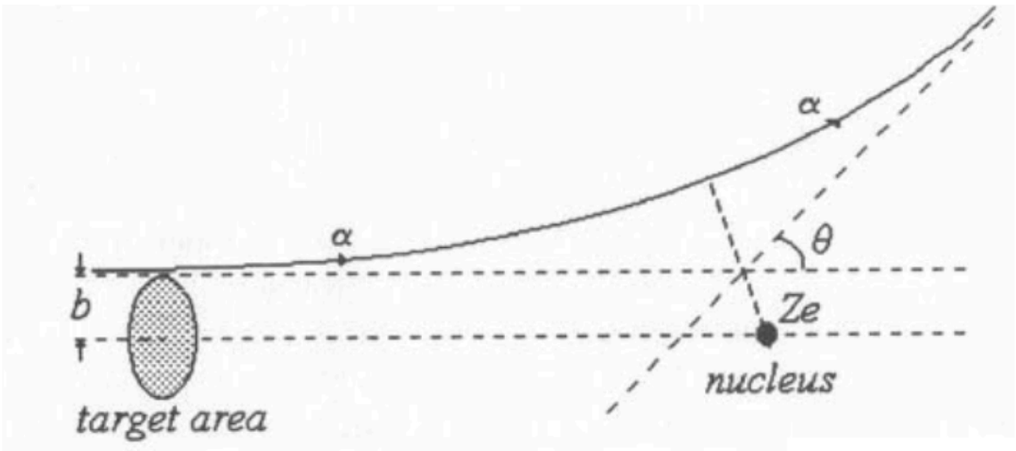
The repulsive force, F_E, between the positive a particle and the positive charge, +Ze, carried by the nucleus is given by Coulomb's Law, and in close encounters this force is very great (Fig 3.4b).

The small angle deflections, which are the vast majority, result from relatively distant encounters between a particles and nuclei, in regions where the electric field and the concomitant repulsive force between the particles is weak.

Rutherford showed, that under the influence of a Coulombic force, the path of an a particle passing by a nucleus would be a hyperbola, the nucleus being situated at its external focus.

The angle of deflection of the particle, θ, is that between the two asymptotes of the hyperbola (Fig 3.6).

Fig 3.6 The path of an α particle as it passes close to a nucleus that carries a nuclear charge of +Ze. The imaginary target area, πb^2 , is that into which the particle must be aimed in order that its deflection be at least through the angle θ.



Every a particle, that is directed at a point within a certain target area around the nucleus will be deflected through an angle equal to or grater than a particular value of θ.

The target area, πb^2 , that corresponds to a particular angle of deflection is called the cross-section for that deflection; b is called its *impact parameter*.

In order to incur a large deflection, an a particle must approach the nucleus with an impact parameter of 10^{-13}m or less.

Rutherford predicted, that the fraction of the incident a particles deflected through any given angle, θ, on traversing the metal foil, would be proportional to $\text{cosec}^4(\theta/2)$.

This was confirmed in a further series of experiments conducted by Geiger and Marsden in 1913.

Rutherford drew attention to the importance of the magnitude of the nuclear charge in the determination of the properties of an atom.

However, though the 'primitive atom' of negative electricity -the electron-had been discovered some fifteen years previously, the particle that carried the natural unit of positive electricity, and which must be a constituent of every nucleus, had not yet been identified.

The identification of the 'primitive atom' of positive electricity was first made by Rutherford. In a suitable discharge tube, positive (canal) rays can be produced that comprise simply hydrogen atoms that have lost their electrons.

No positively charged atom lighter than these had ever been detected.

Furthermore, they never carry more than a single natural unit of positive charge, +e. Rutherford reasoned that the positive particles in these rays must be the nuclei of hydrogen atoms.

He identified these particles-hydrogen nuclei-as the carriers of the natural unit of positive electricity and called them protons.

Thus a neutral hydrogen atom comprises a single proton and a single electron.

To summarize Rutherford's model of the atom:

1. **Atoms comprise a very small and extremely dense nucleus that contains the entire positive charge, $+Ze$, of the atom and virtually all of its mass.**
2. **Z electrons, each of a charge e^-, surround the nucleus at a relatively great distance from it, determining its diameter; together with the nucleus they comprise the neutral atom.**

Part 3.2 The Bohr Model of the Atom

Rutherford was not the first to propose a nuclear structure for the atom.

The idea had been put forward previously in 1903 by Nagaoka.

He had suggested that the atom consists of a massive positive center surrounded by rings of orbiting electrons, like the rings of tiny particles around the planet Saturn.

Such systems had been studied some thirty years earlier by Maxwell who had shown that, in classical mechanics, a condition for the dynamic stability of such systems of rotating particles is that only forces of attraction act between its components.

The only forces acting in the Saturnic system are attractive gravitational forces.

However, the orbiting electrons in Nagaoka's atom would repel one another.

Nagaoka recognized this but argued that it was the small oscillations of the electrons induced by their instabilities that gave rise to the radiation emitted by atoms.

He also suggested that the α and β particles emitted by radioactive atoms were the result of more violent instabilities.

His ideas were not found sufficiently convincing at the time.

A further problem arises from the electromagnetic tenet that all accelerated charges radiate energy in the form of electromagnetic waves.

The orbiting electrons are accelerated and so should radiate electromagnetic waves, until, having lost all their energy, they collapse onto the nucleus (Fig 3.7).

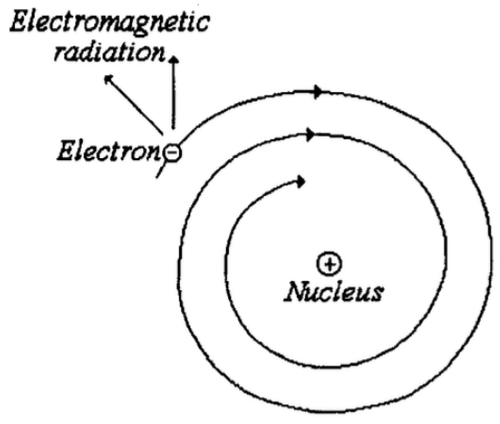


Fig 3.7 The spiral motion of an electron around the nucleus predicted by classical considerations. The circulating electron radiates electromagnetic waves and loses energy. As a result, it 'falls' towards the positive nucleus. Starting from a distance of 1 nm, it would crash onto the nucleus after just 10^{-12}s.

Thus, even a hydrogen atom, with its single electron, should be prohibited from having a nuclear configuration.

Rutherford was not unaware of these problems.

He understood that the nuclear model he had proposed begged the question of the stability of the atomic electrons; the simplistic analogy with the solar planetary system was no solution.

Furthermore, a comprehensive atomic theory could not ignore the incredibly immense density, ~10^{17}kg/m^3, that the nuclear model predicts for the nucleus; it too must be explained.

The first solution to be proposed to the problem of the stability of the electrons in a nuclear atom was that put forward by Bohr in 1913.

Bohr avoided the difficulties presented by classical electromagnetic theory by simply denying its sole or absolute validity in atomic systems.

Instead, he adopted Planck's quantum hypothesis and postulated that an atom emits or absorbs energy in discrete quanta, i.e., that the emission and absorption of energy by an atom is not the continuous process enjoined by classical electromagnetism but is a discrete quantum process.

Bohr then went on to argue that when an atom is not actually emitting or absorbing energy, which is most of the time, it is in a non-radiating stable state which he designated a *stationary state*.

Each of these stationary states is characterized by a specific amount of energy called its *energy level*.

Bohr's third premise was that the emission and absorption of radiation involves transitions between the stationary states.

The quantum of radiation energy -the photon- that an atom emits or absorbs in a transition between two stationary states being the amount of energy specified by the difference in their energy levels.

In distinguishing between the non-radiating stationary states and the radiative transitions between them, Bohr assumed

> *that the dynamical equilibrium of the systems [the atoms] in the stationary states can be discussed by help of the ordinary mechanics, while the passing of the systems between different stationary states cannot be treated on that basis … the latter process is the one given by Planck's theory.*

Bohr proposed that whereas transitions between stationary states are governed by quantum considerations, the stationary states could be described classically.

Thus, when the atom is in a stationary state, specific trajectories -stationary orbits- whose parameters are determined by classical considerations, can be assigned to the electrons.

Although now superseded by the quantum mechanical model of the atom, Bohr's model marked a major breakthrough in our understanding of atomic structure.

It was the first model to offer a comprehensive explanation of how atomic structure determines the nature of the radiation emitted and absorbed by an atom.

We shall illustrate this by showing how the model dealt with the following three phenomena:

1. **The spectra of the hydrogen atom, H, and of hydrogen-like atoms, such as He^+ and Li^{++}, each of which comprises a nucleus and a single electron. These spectra are particularly well-defined in the optical region. This enabled precise measurements to be made which resulted in the discovery of a number of empirical relationships.**
2. **The Zeeman Effect: the splitting of spectral lines in a magnetic field.**
3. **The characteristic radiation emitted from X-ray tubes: Moseley's experiment.**

3.2.1 The Hydrogen Atom

Bohr began his study of hydrogen and hydrogen-like atoms by presupposing that the interaction between the atom's negative electron and its positive nucleus is Coulombic in nature.

Thus, the force of attraction, F_E, that acts between an electron and a nucleus carrying a charge +Ze is given by

$$F_E = -kZe^2/r^2 \tag{3.1}$$

where r is the distance between them and k is the electrostatic constant of proportionality

Classically, when two particles are bound by a central force such as the electrostatic force of attraction that acts between two oppositely charged particles, the smaller particle orbits the larger one in an elliptical path.

Applying this principle to the structure of hydrogen or hydrogen-like atoms, Bohr postulated that in their non-radiating stationary states, the tiny electron revolves around the massive nucleus in an elliptic or, in the simplest case, a circular orbit.

In the case of a circular orbit of radius R, the electron's potential energy, U, and kinetic energy, K, are given, respectively, by the equations:

$$U = -kZe^2/R \tag{3.2}$$

$$K = m_e v^2/2 = kZe^2/2R \tag{3.3}$$

where m_e is the mass of the electron and v its orbital velocity.

Accordingly, the total energy, E, of the orbiting electron is:

$$E = U + K = -kZe^2/2R \tag{3.4}$$

The orbiting electron's total energy is negative because it is bound to the nucleus, i.e., energy must be invested to release it.

Having established these classical expressions for the electron's energy in a circular stationary orbit, Bohr now made the following *ad hoc* assumptions concerning the manner in which energy might be released during the formation of a hydrogen atom from a free proton and a free electron, i.e., as the electron spirals down from infinity and settles into a stationary orbit around the proton.

> *Let us assume that, during the binding of the electron a homogeneous [monochromatic] radiation is emitted of frequency φ, equal to half the frequency of revolution of the electron in its final orbit; then, from Planck's theory, we might expect that the amount of energy emitted by the process is equal to nhφ whew h is PIanck's constant and n an entire [integer] number…. the assumption concerning the frequency of the radiation suggests itself since the frequency of revolution of the electron at the beginning of the emission is zero.*

Bohr supposed that during the atom's formation, n photons of a hypothetical radiation of frequency φ would be emitted, thereby releasing an amount of energy $\Delta E = nh\varphi$.

He further suggested that the frequency, φ, of the hypothetical monochromatic radiation could be assumed to be equal to the mean of the initial and final frequencies of the electron's spiral.

The electron's initial frequency of revolution at infinity is zero.

Hence, if the frequency of the stationary orbit into which it finally settles is f; the frequency of the hypothetical radiation will be $\varphi = (0 + f)/2 = f/2$.

The frequency of revolution, f; of an electron circulating at a constant velocity v in an orbit of radius R is $f = v/2\pi R$ from which:

$$\varphi = {}^{v}/_{4\pi R} \tag{3.5}$$

Thus, the total energy released during the formation of the atom would be:

$$\Delta E = nh\varphi = \frac{nh\nu}{4\pi R} \tag{3.6}$$

Assuming that the electron started out with zero energy at infinity, its energy in the stationary state will be

$$E = 0 - \Delta E = -\frac{nh\nu}{4\pi R} \quad (n = 1,2,3,\ldots) \tag{3.7}$$

where n is the number of photons of the hypothetical radiation emitted during the atom's formation.

Based on his premise that the stationary orbits could be described classically, Bohr equated this expression with that - equation (3.4)- given by classical considerations for the electron's total energy

$$E = -kZe^2/2R = -nh\nu/4\pi R \tag{3.8}$$

Finally, comparing the classical expression for the orbiting electron's kinetic energy,

$$K = m_e v^2/2 = kZe^2/2R$$

with that for its total energy

$$E = -kZe^2/2R$$

shows that in a circular orbit the electron's kinetic energy is equal to the magnitude of its total energy, i.e., K =|E|. Hence,

$$m_e v^2/2 = nh\nu/4\pi R \tag{3.9a}$$

from which

$$m_e vR = n \cdot \frac{h}{2\pi} = n\hbar \tag{3.9b}$$

The quantity m_eVR, that appears in this equation, is the orbital angular momentum, L, of the electron in the stationary orbit.

Thus, equation (3.9b) can be re-written as

$$L = n\hbar \quad (n = 1,2,3,\ldots) \tag{3.10}$$

This simple equation expresses the key concept introduced into physics by Bohr's atomic theory, namely, that an atomic electron's orbital angular momentum is *quantized*, it can only take certain values which are all integer multiples of $\hbar$.

The integer n, called the *principal quantum number*, designates the electron's orbit.

The stationary orbit corresponding to n = 1 is called the *ground state* (Fig 3.8).

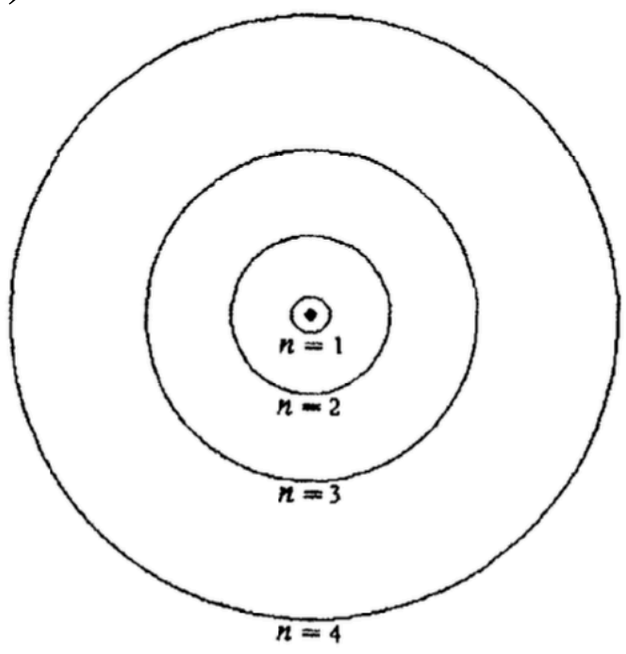


Fig 3.8 The stationary electronic orbits predicted by Bohr's model of the hydrogen atom. The circles represent the orbits corresponding to the principal quantum numbers 1, 2, 3 and 4.

Having derived this result from the quantum hypothesis, Bohr went on to suggest that it was a general rule that applied to all electrons moving in circular stationary orbits, irrespective of the nuclear charge.

concept of angular momentum quantization subsequently assumed the status of an axiom, underivable from any deeper law, depending for its validity on the consistent agreement between its corollaries and experimental results.

It is the usual starting point for derivations of Bohr's atomic model.

The radii, R_n, of the stationary circular orbits in hydrogen and hydrogen-like atoms, such as He^+ and Li^{++}, are obtained by substituting the expression for the electron's velocity, $v = n\hbar/m_eR$ obtained from equation (3.09), in the classical expression for its kinetic energy, $K = m_ev^2/2 = kZe^2/2R$.

This gives:

$$R_n = \frac{n^2\hbar^2}{m_e kZe^2} \qquad (n = 1,2,3,\ldots) \tag{3.11}$$

The radius of the ground state (n = 1) in the hydrogen atom is usually designated a_0 and termed the *Bohr radius*: $a_0 = 5.29\text{x}10^{-11}\text{m}$.

The radius of the ground state orbit in He+(Z=2) is $a_0/2$ and in Li^{++} it is $a_0/3$.

The principle energy levels, E_n, in hydrogen and hydrogen-like atoms are obtained by substituting the radii of the stationary orbits in the expression for the electron's total energy, $E = - kZe^2/2R$.

This gives:

$$E_n = -\frac{k^2 Z^2 m_e e^4}{2n^2 \hbar^2} \quad (n = 1, 2, 3, \ldots) \tag{3.12}$$

The energy levels in H, He^+ and Li^{++} are shown diagramatically in Fig 3.9.

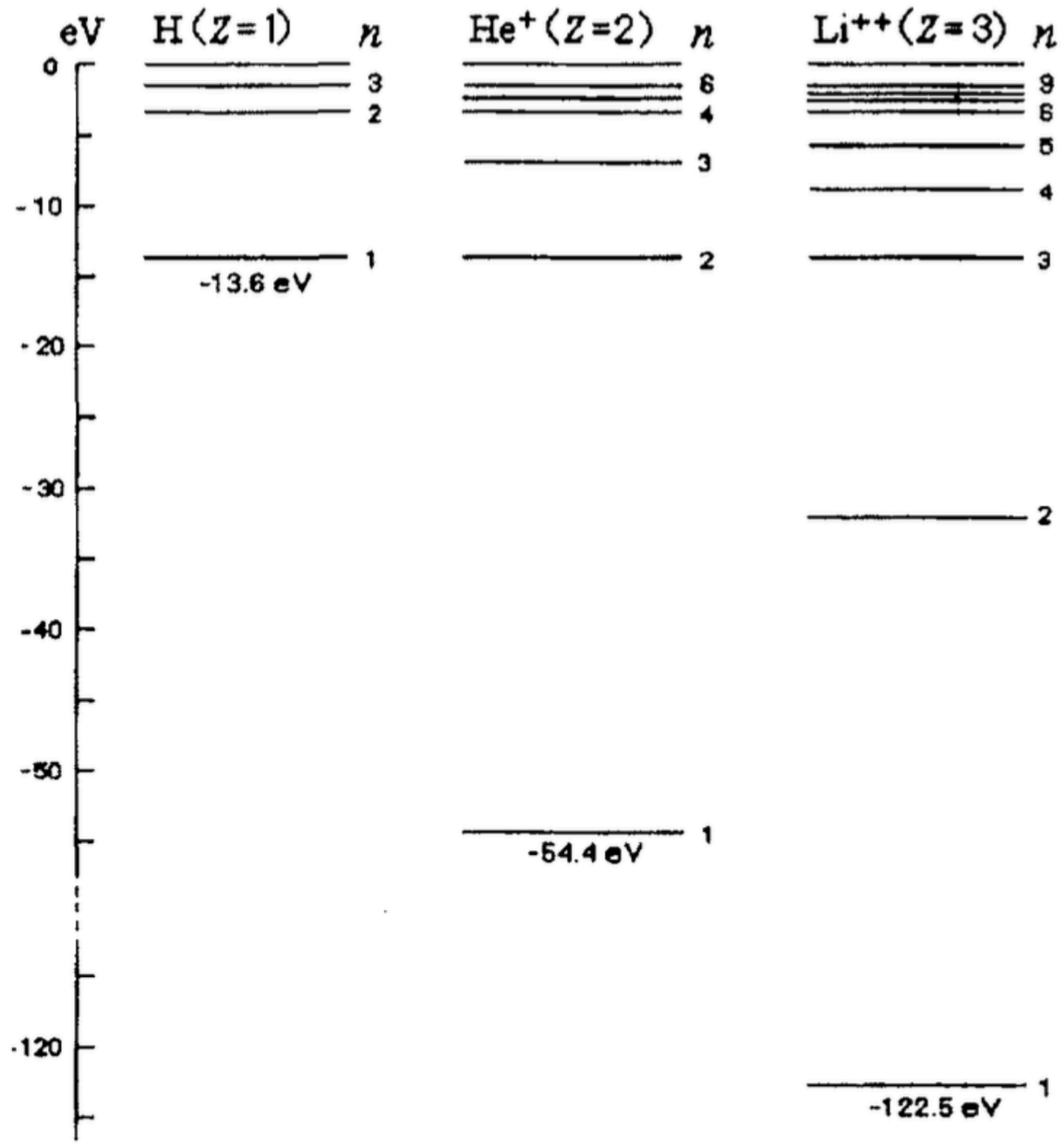


Fig 3.9 The energy level diagrams of H, He^+ and Li^{++}. As the principal quantum number, n, increases, the spacing between the energy levels decreases and the atomic electron's total energy approaches zero.

In terms of the nuclear charge, Z, the energy of the ground state (n = 1) in each of the three atoms, H, He^+ and Li^{++}, is given by:

$$E_1 = 13.5Z^2\text{eV} \tag{3.13}$$

This defines the minimum amount of energy -the ionization energy- that must be invested in order to remove the orbiting electron from a hydrogen or hydrogen-type atom in its ground state.

Thus, Bohr's model predicts that the ionization energy of hydrogen (Z = 1) should be 13.6eV, that of He^+ (Z = 2) 13.6x 22 = 54.4eV, and that of Li^{++} (Z = 3) 13.6x 32 = 122.4eV .

These figures are all confirmed experimentally.

Having derived a general formula for the energy levels in hydrogen and hydrogen-like atoms, Bohr went on to show how, when taken together with his postulate that transitions between stationary states are accompanied by the emission or absorption of a photon, it could account for the regularities between the wavelengths and frequencies of the lines in the hydrogen emission spectrum that had been recognized for some time and for which no satisfactory explanation had yet been found.

Chief amongst these empirical protocols were the *Ritz Combination Rule* and the *Rydberg Equation*.

As we showed in Part 2.1, the combination rule formulated by Ritz follows directly from energy conservation and Bohr's equation,

$$f_{ij} = \frac{E_i - E_j}{h}$$

for the frequency of the photon emitted or absorbed in a transition between the energy levels E_i and E_f.

The empirical equation that Rydberg had formulated for the wavelengths of the lines in the emission spectrum of hydrogen was

$$\frac{1}{\lambda} = R_0\left(\frac{1}{n_j^2} - \frac{1}{n_i^2}\right) \tag{3.14}$$

where R_0 is a constant known as *Rydberg 's constant for hydrogen*, $n_j = 1,2,3,,$ .. and $n_i = n_j + 1, n_j + 2, n_j + 3,$ ...

The experimental value found for the constant R_0 was$1.09\text{x}10^7 m^{-1}$.

A special case of this general relationship had been discovered earlier Balmer in 1885 when he noted that the wavelengths of the lines in the hydrogen spectrum in the visible region could all be expressed by the simple formula

$$\frac{1}{\lambda} = B\left(\frac{1}{2^2} - \frac{1}{n_i^2}\right) \qquad n_i = 3,4,5,\ldots \tag{3.14a}$$

where B is a constant.

Bohr showed that Rydberg's equation and the value of the constant R_0 could be directly derived from his theoretical model, as follows. From the expression he had derived for the energy level of a stationary state-equation (3.12) the difference between the energy level, E_i, designated by the principal quantum number n_i and the energy level, E_j, designated by n_j is

$$E_i - E_j = -\frac{k^2 m_e e^4}{2\hbar^2}\left(\frac{1}{n_j^2} - \frac{1}{n_i^2}\right) \tag{3.15}$$

A transition between these states would be accompanied by the emission or absorption of a photon of frequency f_{ij} as given by Bohr's equation:

$$f_{ij} = \frac{E_i - E_j}{h} = \frac{k^2 Z^2 m_e e^4}{2\hbar^2 h}\left(\frac{1}{n_j^2} - \frac{1}{n_i^2}\right) \tag{3.16}$$

Making the substitutions $c = \lambda f$ and Z = 1 gives for the hydrogen atom:

$$\frac{1}{\lambda_{ij}} = \frac{k^2 m_e e^4}{2\hbar^2 hc}\left(\frac{1}{n_j^2} - \frac{1}{n_i^2}\right) \tag{3.17}$$

The similarity between the form of this equation and that of the empirical expression formulated by Rydberg -equation (3.14)- is evident.

Furthermore, the agreement is quantitative as well as qualitative.

The calculated SI value of the quotient

$$\frac{k^2 m_e e^4}{2\hbar^2 hc}$$

outside the brackets in this equation is $1.093 \times 10^7 m^{-1}$, almost exactly the same as the experimental value of Rydberg's constant for hydrogen.

Summarizing his argument, Bohr wrote:

> *We see that this expression accounts for the law connecting the lines In the spectrum of Hydrogen. If we put $n_j = 2$ and let n_i vary, we get the ordinary Balmer series. If we put $n_j = 3$, we get the series in the ultraed (infra-red] observed by Paschen ... If we put $n_i = 1$ and $n_j = 4,5,...$, we get series respectively in the extreme ultraviolet and the extreme ultrared [infra-md], which are not observed but the existence of which may be expected.*

The series for which $n_i = 1$ and $n_j = 4$ were discovered soon afterwards.

The former is now known as the Lyman series and the latter as the Brackett series.

The transitions associated with the four series are shown in Fig 3.10.

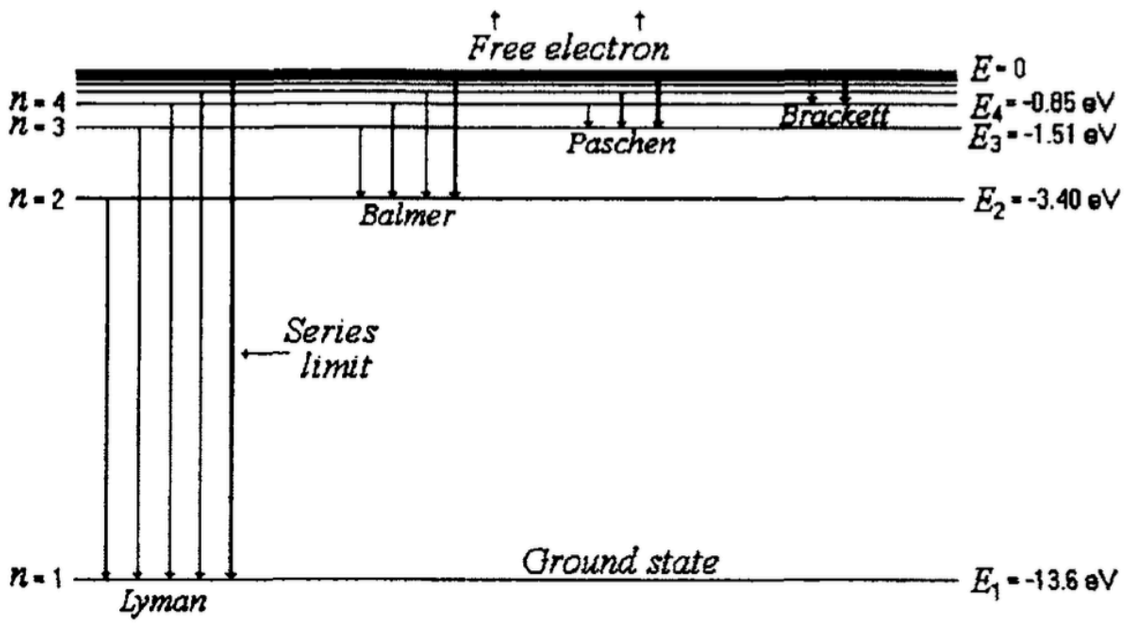


Fig 3.10 The energy levels and transitions that are responsible for the emission spectrum of a hydrogen atom. The spectral lines in the Lyman series correspond to transitions from higher levels to the ground state: $n_j = 1$. Those in the Balmer series to transitions from higher levels down to the $n_j = 2$ level, those in the Paschen series down to the $n_j = 3$ level and those in the Brackett series down to the $n_j = 4$ level. When $E \geq 0$ the electron is free and no longer in orbit around the nucleus

In his original calculation, Bohr had assumed, as we have done so far, that the mass of the electron orbiting the nucleus was its rest mass, m_e.

This would be correct if the hydrogen nucleus was of infinite mass and were it fixed at the point in space around which the electron rotates in its stationary circular orbits.

But the mass of the nucleus, M, is not infinite and so the nucleus and the electron actually revolve around that such a system is equivalent to their common center of mass.

It can be shown, that such a system is equivalent to one in which a single particle of *reduced mass*, m', given by the formula

$$m' = \frac{m_e}{1 + m_e/M} \tag{3.18}$$

revolves around the larger mass.

In the case of a normal hydrogen atom $m' = 0.99945 m_e$.

Though not large, differences due to this effect are detectable at high resolution.

For example, the isotope of hydrogen, deuterium, whose mass is almost exactly twice that of a hydrogen atom, was first detected spectroscopically in 1932, as a result of the difference between its reduced mass, $m' = 0.99973 m_e$, and that of normal hydrogen, $m' = 0.99945 m_e$.

Because of this difference, the spectral line corresponding to the transition from n = 3 to n = 2 that occurs at a wavelength of 656.3nm in the spectrum of hydrogen, occurs at 656.1nm in the deuterium spectrum.

In a sample of naturally occurring hydrogen, the intensity of the deuterium spectral lines is about1/6000 that of the corresponding normal hydrogen lines, a ratio which reflects their relative abundance.

Following Bohr's initial analysis of the circular stationary orbits, Sommerfeld went on to investigate the elliptical orbits postulated by the model.

Their description required the separation of the principal quantum number, n, into a radial quantum number, n', and an azimuthal quantum number, k, such that $n = n' + k$. It then became apparent that it is actually the azimuthal quantum number, k, and not the principal quantum number, n, that fixes the electron's total orbital angular momentum:

$$L = k\hbar \quad (k = 1, 2, 3 \ldots, n) \tag{3.19}$$

In the case of a circular orbit k = n, and so the need for more than one quantum number did not arise in Bohr's original derivation.

At high resolution, the major spectral lines, such as those in the Balmer series, are seen to comprise a number of finer lines.

Sommerfeld explained that this fine structure within the major spectral lines was the result of relativistic effects, such as the periodic variations in the electron's mass as its velocity varied in its elliptical orbit around the nucleus.

He suggested, that these effects would cause a precessional motion of the axis of the elliptical orbits similar to that of the planet mercury.

This precessional motion has a small energy associated with it, whose effect is to split the principal energy levels, E_n, set by the quantum number n into as many sub-levels.

The fine structure is associated with transitions from the sub-levels, k = 1,2, ...,n_2, in an upper state, to those, k = 1,2,.. .,n_1, in a lower state (Fig 3.11).

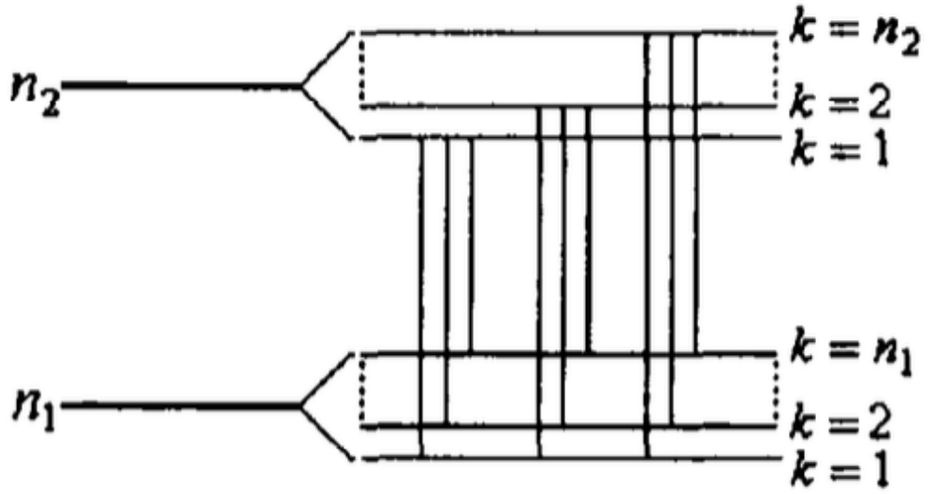


Fig 3.11 An idealized diagram of the possible transitions from sub-levels in an upper state, n_2, to sublevels in a lower state, n_1 . The existence of the sub-levels is indicated by the fine structure of narrow lines, that is observed within the broader principal spectral line corresponding to the transition from n_2 to n_1. The finer spectral lines correspond to transitions from sublevels in the upper state to sublevels in the lower state.

Since in general k = 1,2,3,. ..,n , the higher the principal quantum number, n, the more k values there are and the greater the number of different transitions that can be imagined.

It might therefore be expected that a correspondingly large number of fine structure lines would be observed.

However, this is not so.

The only transitions that actually show up as spectral lines are those for which $\Delta k = \pm 1$.

Such expressions, that show in terms of the quantum numbers which transitions are allowed, are known as *selection rules*.

They are of paramount importance in the interpretation of spectra and indicate that factors other than just energy conservation are involved.

3.2.2 The Zeeman Effect Space Quantization

Based on Maxwell's theory that electromagnetic waves are generated by accelerated charges, Lorentz had predicted that atoms must contain electrical charges.

This being so, emission spectra should be affectedby a magnetic field; it should alter the motion of the charges.

In 1896, a year before the discovery of the electron, Zeeman confirmed this prediction when he discovered the splitting of the spectral lines of the light emitted from a source situated in a magnetic field; two or three adjacent lines are observed where, in the absence of the magnetic field, just one is seen.

This phenomenon, called the *Zeeman Effect*, was first explained classically by Lorentz.

He showed that a magnetic field may increase the frequency of one component of an electric oscillator's vibrations by exactly the same amount as it decreases the opposite component.

Consequently, in the presence of a magnetic field, an atomic oscillator may emit more than one frequency and so more than one spectral line may appear.

The Bohr model offers a different explanation.

It attributes the Zeeman effect to a precessional top-like motion of the normal to the electron's orbital plane relative to the direction of the external magnetic field (Fig 3.13).

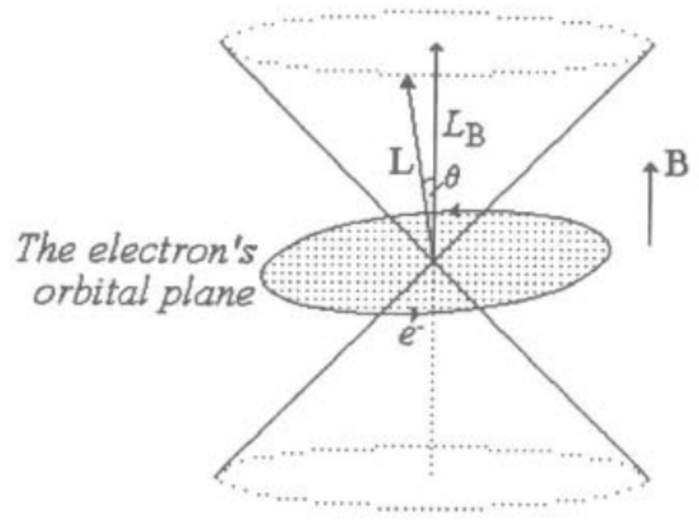


Fig 3.13 The toplike precession of the normal to the plane of the electron's orbit and of its angular momentum vector, **L**, around an axis defined by the direction of the external magnetic field **B**. The normal describes the surface of a cone. L_B is the component of the electron's angular momentum parallel to the magnetic field.

The angular momentum associated with this motion is quantized and the additional spectral lines arise from transitions between the energy sub-levels that accompany this.

An orbiting atomic electron possesses a magnetic dipole moment, μ, of magnitude

$$\mu = -eL/2m_e \tag{3.20}$$

which it confers on the atom.

Considered as a vector, the magnetic dipole, μ, is parallel to the angular momentum vector, **L**, i.e., normal to the orbital plane.

In a magnetic field, **B**, a torque, $\tau = \mu$ x **B**, acts on the magnetic moment.

This causes the atom's magnetic moment to precess around the direction of the field, like the motion of a spinning top in a gravitational field.

The frequency, f_L, of this precession, called the *Larmor Frequency*, is given by:

$$f_L = \frac{\mu B}{2\pi L} \tag{3.21}$$

Classically, the normal to the plane of the electron's orbit can precess at any angle relative to the direction of the magnetic field.

In contrast, on the Bohr model, the sub-levels responsible for the Zeeman splitting arise from the introduction of a further quantization condition that has the effect of restricting the orbital plane's precession to certain specific angles relative to the direction of the magnetic field.

This additional, apparently arbitrary, quantization condition requires that the component, L_B, of the electron's angular momentum parallel to the direction of the external magnetic field be an integer multiple of $\hbar$, i.e.,

$$L_B = m\hbar \quad (m = -k, -k+1, -k+2 \ldots, +k) \tag{3.22}$$

It follows, that the normal to the electron's orbital plane and the total orbital angular momentum vector, **L**, can only take up specific inclinations, θ_m, relative to the direction of the external magnetic field, i.e., those that satisfy the condition:

$$\cos\theta_m = \frac{L_B}{L} = \frac{m\hbar}{L} \tag{3.23}$$

The phenomenon is called the Quantization of Direction or Space Quantization and its existence was first explicitIy demonstrated by Stern and Gerlach in an experiment carried out in 1921.

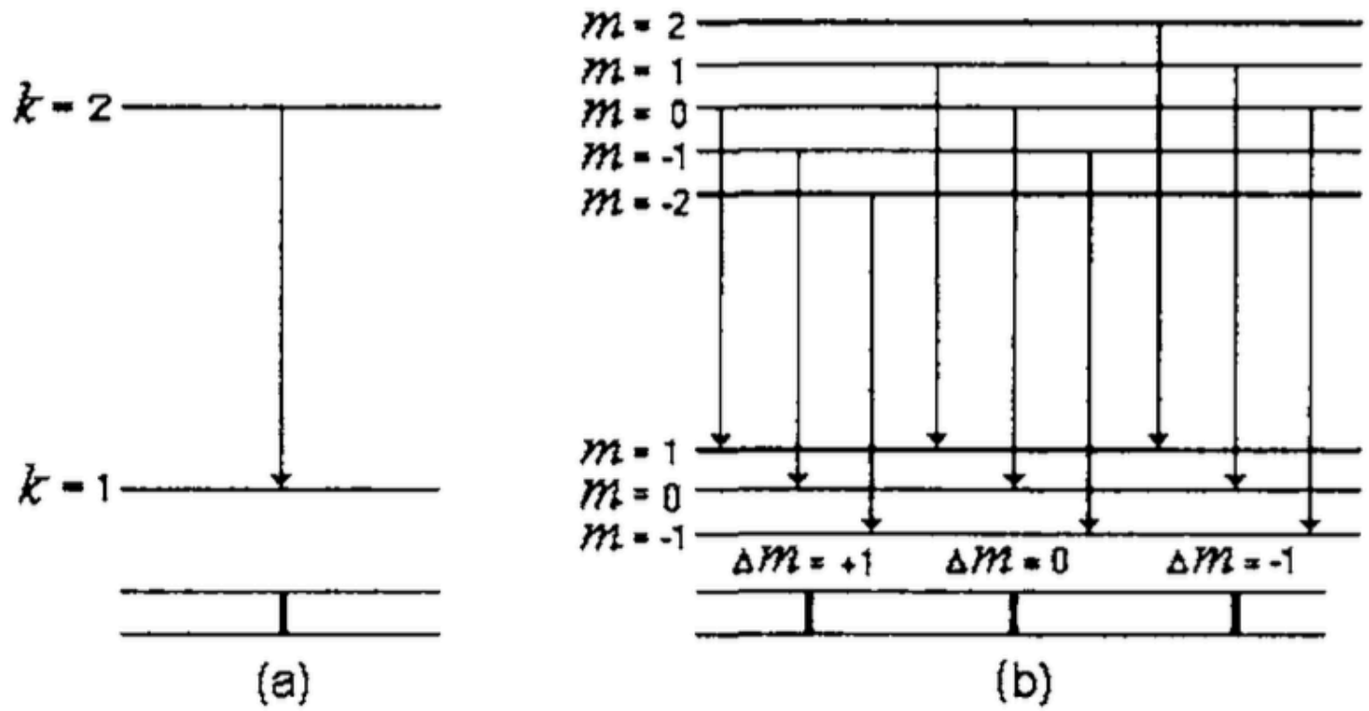


Fig 3.14 The Zeeman effect. (a) The single spectral line observed in the absence of a magnetic field that corresponds to the transition from the k = 2 to the k = 1 state. (b) The three lines that appear in the presence of a magnetic field as a result of the splitting of the levels into sub-levels. Each line corresponds to a different value of Δm.

The energy, E_μ, of a magnetic dipole moment, μ, situated in a magnetic field, B, is given by:

$$E_\mu = -\boldsymbol{\mu}\cdot\mathbf{B} = \mu B\cos\theta \tag{3.24}$$

nowhere θ is the angle between the direction of the magnetic dipole and that of the field. Substituting the space quantization provision,

$$\cos\theta_m = \frac{m\hbar}{L} \tag{3.25}$$

in this equation gives the following simple expression for the energy sub-levels responsible for the Zeeman effect:

$$E_\mu = {}^{\mu B m\hbar}/_{L} \quad (m = -k, -k+1, -k+2 \ldots, +k) \tag{3.26}$$

In terms of the dipole's Larmor frequency,

$$f_L = \frac{\mu B}{2\pi L}$$

this reduces to:

$$E_\mu = mhf_L \quad (m = -k, -k+1, -k+2 \ldots, +k) \tag{3.27}$$

According to this model, the Zeeman splitting of the spectral lines occurs when transitions between the k states are accompanied by transitions between the sub-levels defined by the additional quantum number m.

In each such transition a photon of energy hf_L will be absorbed or emitted.

Although every k state is split into (2k + 1) sublevels, it transpires, that only transitions that obey the selection rules, $\Delta m = \pm 1$ or $\Delta m = 0$, show up in observed spectra (Fig 3.14).

An important corollary of the postulate of angular momentum quantization is that there exists a natural unit of atomic magnetic moment.

The absolute value of this unit, which is designated μ_B, is obtained by substituting $L = \hbar$ in equation (3.20):

$$\mu_B = e\hbar/2m_e = 9.27 \cdot 10^{-24}\,\mathrm{J \cdot T^{-1}} \tag{3.28}$$

This quantity is called the *Bohr Magneton*.

3.2.3 Moseley's Experiment

The high intensity penetrating radiation emitted by X-ray tubes, characteristic of the metal from which the target anode is made, was first discovered by Barkla.

He found that when the tubes were operated at higher potentials, series of high intensity peaks, each of a specific wavelength, were superimposed on the spectrum of the continuous brehmstrahlung radiation (Fig 3.15).

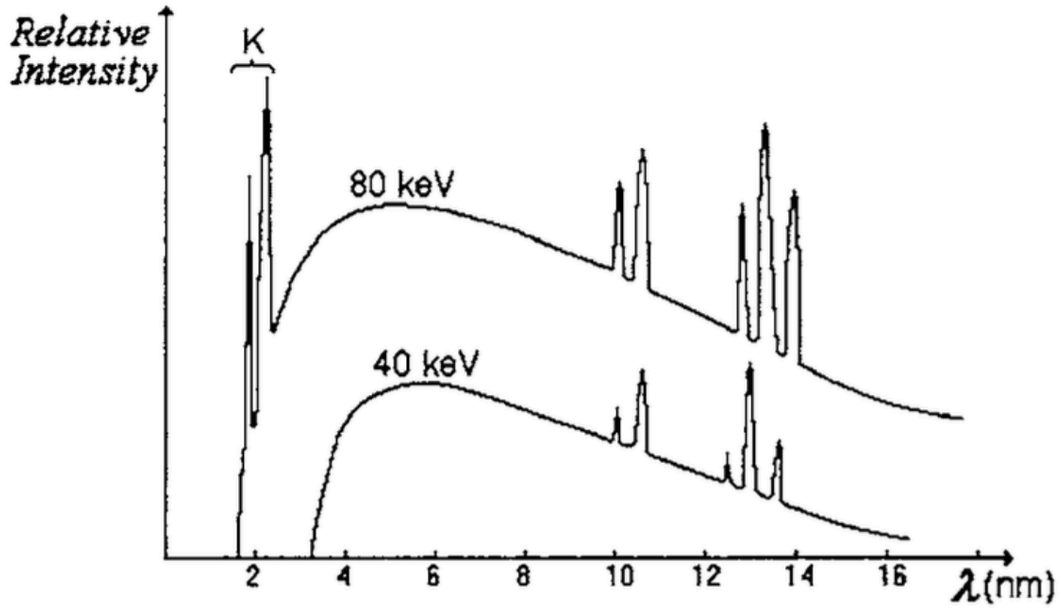


Fig 3.15 The spectrum of the radiation emitted from an X-ray tube operated at potentials of 40 keV and 80 keV. The more penetrating K series only appears in the spectrum at the higher potential because the 40 keV electrons have insufficient energy to dislodge electrons from the K shell.

The phenomenon is analogous to the atomic line spectra seen in the visual region of the electromagnetic spectrum.

Changing the metal or element from which the target anode in the X-ray tube is made alters the wavelengths at which the high intensity peaks occur.

The most penetrating series in an element's characteristic X-ray spectrum is called the K series; the second is called the L series; the third the M series and so on.

At the time of their discovery, the available experimental techniques did not enable an accurate determination of the wavelengths or frequencies of the characteristic X-rays to be made.

However, after Bragg had shown how such measurements could be made with the help of crystals, Moseley decided to use this technique in order to carry out a systematic examination of the characteristic radiation of as many elements as possible.

He examined the X-ray spectra of the 38 elements from aluminium (Al) to gold (Au). As regards 15 of these elements, he studied just the K series; regarding another 17, just the L series; as to the remaining 6 elements, both series.

He recorded the spectra on photographic plates.

Moseley discovered the following simple empirical relationship, illustrated in Fig 3.16, between the frequencies, f_j of the lines in each series and the ordinal number, N , of the element's position in the periodic table (starting from hydrogen):

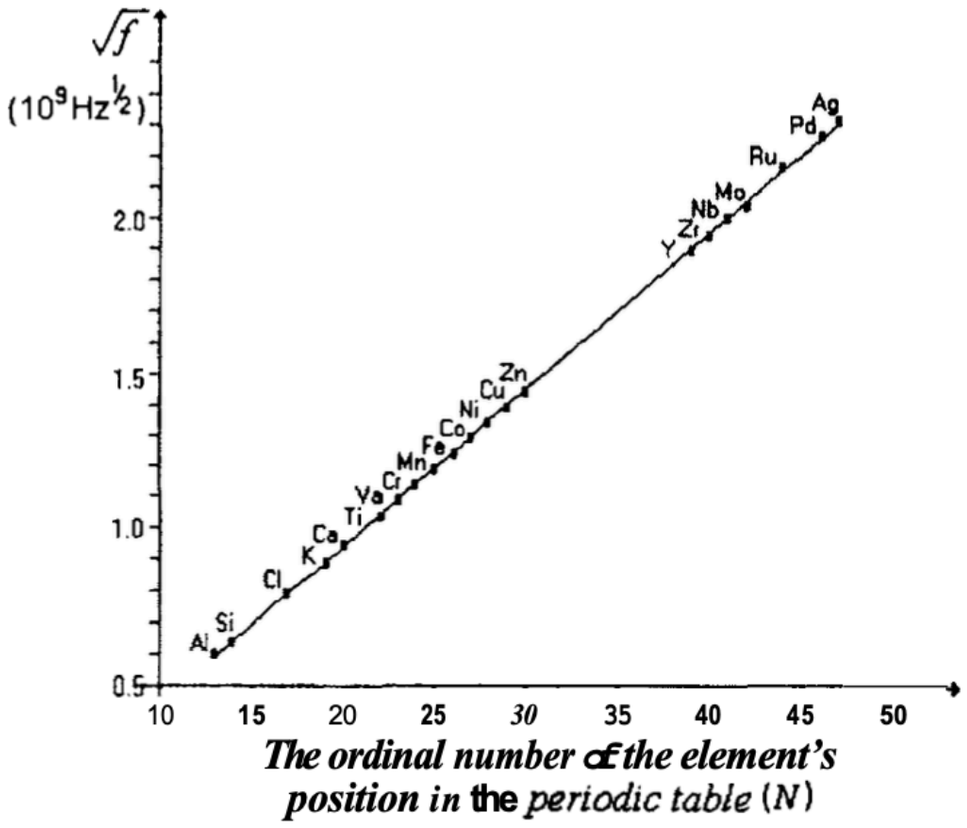


Fig 3.16 The square root, of the frequency of an element's $\sqrt{f}$, line as a function of the ordinal number, N, of its position in the periodic table.

Moseley formed the opinion that some physical attribute of the atom must increase by a regular fixed amount, from one element to the next, rising through the periodic table.

He postulated that this could only be the atom's nuclear charge.

According to this hypothesis, the number N, that is the element's ordinal position in the periodic table, is equal to the number of natural units of positive electricity carried by the nuclei of the element, i.e., N=Z.

The number Z is now called the *atomic number* of the element; it is equal to the number of protons in the element's nuclei.

Prior to Moseley's investigation, the elements were arrayed in the periodic table in the ascending order of their atomic weights and on the basis of their chemical properties.

As a result of Moseley's researches, which provided the first direct means of determining an element's atomic number, inaccuracies in the periodic table were discovered and corrected.

For example, the positions of the transition metals cobalt (Z = 27) and nickel (Z = 28), that had been previously determined by the ascending order of their atomic weights, Ni = 58.71 and Co = 58.93, were switched.

Similarly, empty positions were revealed in the table, corresponding to the yet undiscovered elements of atomic number 43, 61, 72 and 75.

The origin of the characteristic X-ray radiation is readily explained by the Bohr model of the atom.

Let us assume, that the electrons orbiting the nucleus in multielectron atoms are arranged in shells, each electron having its specific slot in a shell.

We will designate the innermost shell as the *K shell*, it corresponds to the principal quantum number n = 1.

The second shell is designated the *L shell*; it corresponds to the principal quantum number n = 2.

The third shell is designated the *M shell*; it corresponds to the principal quantum number n = 3, and so on.

According to this model, all the electrons in a particular shell have the same energy and the closer a shell is to the nucleus the greater the energy binding its electrons to the atom.

The electrons in the K shell are all in the energy level E_1, those in the L shell in the level E_2, and so on (Fig 3.17).

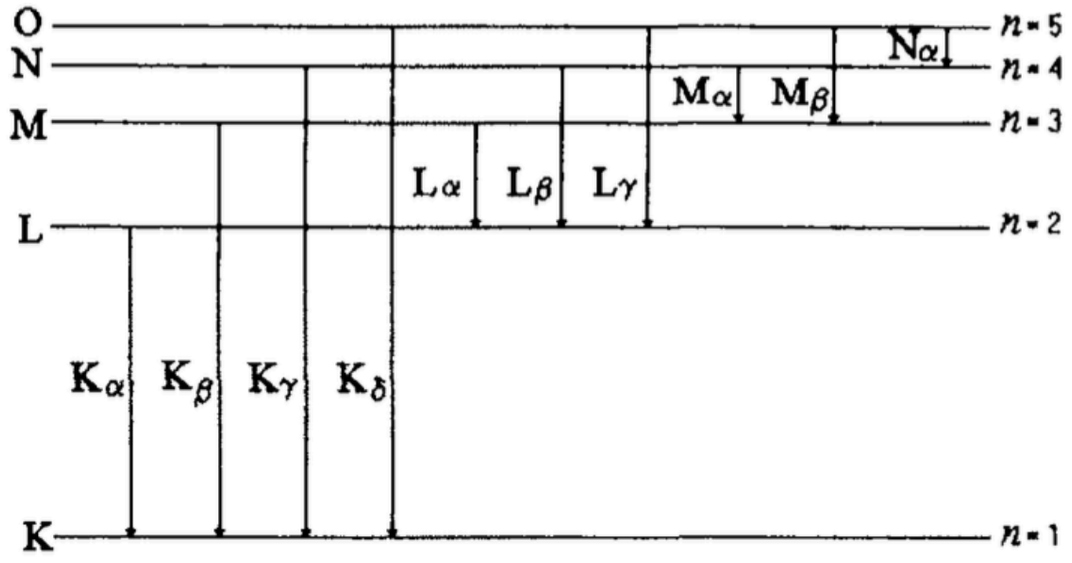


Fig 3.17 The transitions that produce the characteristic X-ray radiations. A 'hole' in an inner shell of the atom's electronic structure is filled by an electron that 'falls' from a higher energy level; the difference in energy appears as a high energy photon.

When an electron, with sufficiently high energy, strikes an atom in the X-ray tube anode, it ejects an electron from one of the atom's inner shells, say the K shell.

This leaves an empty slot or hole in this shell.

One of the electrons from an outer shell, corresponding to an energy level E_n, can 'fall' into this hole, releasing an amount of energy, $E_n - E_l$, equal to the difference between the energy levels of the two shells; this energy is released as an X-ray photon of frequency:

$$f = \frac{E_n - E_1}{h} \tag{3.29}$$

On the basis of this explanation, the K series results from the filling of holes in the K shell (n = l), with the K_α line corresponding to the hole being filled by an electron falling from the L shell; the K_β line to the electron falling from the M shell, and so on.

Similarly, the L series results from the filling of holes in the L shell (n = 2).

The more electrons there are in an atom, the more series there will be in its X-ray spectrum and the more lines there will be in each series.

An atom's electronic energy levels depend on the interaction between its electrons and its nuclear charge, +Ze, and so, since the atoms of each element carry a characteristic nuclear charge, each element exhibits a characteristic X-ray spectrum.

To a close approximation, the frequencies, $f_{K\alpha}$, of the K_α line - the most intense line in an element's K series - are given by the formula

$$f_{K_\alpha} = cR_0(Z-1)^2\left(\frac{1}{1^2}-\frac{1}{2^2}\right) \tag{3.30}$$

It is as though after the removal of an electron from the K shell, the electron in the L shell sees a nuclear charge of (Z - 1).

The frequencies, $f_{L\alpha}$, of the L_α line -the most intense line in the element's L series- are given by the formula:

$$f_{L_\alpha} = cR_0(Z-7.4)^2\left(\frac{1}{2^2}-\frac{1}{3^2}\right) \tag{3.31}$$

The similarity between these formulae and equation (3.16), that gives the lines in the hydrogen spectrum, is apparent.

On account of the relatively large nuclear charge of the anode metals ($Z > 30$), the motion of their innermost electrons is virtually unaffected by the outermost electrons.

The innermost electrons interact almost exclusively with the nucleus, just like the electron in the hydrogen atom.

Consequently, as in Bohr's elucidation of the hydrogen spectrum, the denominators of the fractions in equations (3.30) and (3.31) denote the principal quantum numbers of the energy levels between which the electronic transitions have taken place.

The screening constant, σ, is greater in equation (3.31) because the electrons in the L shell are shielded from the nuclear charge by the innermost electrons in the K shell.

Accurate measurements show, that each principle line in the characteristic X-ray spectrum is in fact composed of a fine structure of very close discrete lines.

For example, the K_α line comprises two lines; it is a *doublet*.

It follows, that when an electron falls from the L shell to the K shell, there are two possible values for the difference, $E_2 - E_1$, between the energy levels of the shells.

This indicates that there are sublevels within the principle energy levels of the atom.

A full explanation of the existence of these sublevels requires the concept of electron spin .

To summarize Bohr's model

1. **Atoms comprise a very small and very dense nucleus that contains their entire positive charge, $+Ze$, and virtually all of their mass.**
2. **In a neutral atom, Z negative electrons revolve around the nucleus in stationary non-radiating orbits: these electrons, together with the nucleus, comprise the neutral atom.**
3. **The interaction between the atomic nucleus and the electrons is Coulombic.**
4. **An atom absorbs and emits energy in accordance with the quantum hypothesis.**
5. **The orbits of the electrons in multi-electron atoms are arranged around the nucleus in shells.**

Though revolutionary and seminal, Bohr 's theory was incomplete.

Like Compton's explanation of the scattering of X-rays by free electrons, it was an amalgam, almost an alchemy, of the quantum hypothesis with classical and relativistic mechanics, that worked in a few simple instances but which had no clear and consistent conceptual basis of its own.

Quantitatively, the model was successful only in the case of the hydrogen atom, the simplest of atoms; its application to more complex systems was never better than semiquantitative.

Furthermore, the model was unable to account for the fact that some of the transitions it forecast didn't show up as spectral lines; nor could it account for the different intensities of the spectral lines that did appear.

And perhaps most serious of all, experiments showed that in its ground state, n = 1, the angular momentum of the electron in the hydrogen atom is zero, and not $\hbar$ as predicted by the axiom of angular momentum quantization.

Moreover, their stability was not the only problem presented by the nuclear atom's electrons.

Parallel to the efforts of the physicists to find an atomic model that would satisfy the requirements of spectroscopy and of the resonance of energy emission and absorption, the chemists were looking for a satisfactory explanation to valency and to the stability of the bonds between the component atoms of molecules and ionic crystals.

Both the physicists and the chemists were dealing with the same atoms, however, the notion of electrons revolving at frequencies of billions of cycles per seconds appeared to preclude their participation "in the process of binding atom to atom in the molecule".

> *The view based on the periodic law and the chemical behavior of the elements leads to the picture of a relatively static atom ... the electron as a whole is regarded as occupying a fixed position in the atom.*

The chemical stability of the non-radioactive elements, i.e., the fact that they do not undergo spontaneous chemical degradation, requires that the atom's electrons, together with its nucleus, constitute an unchanging self-sufficient unit.

Furthermore, molecular architecture-the fixed shape and configuration of molecules-requires that an atom's valence electrons should be 'pointing' in specific directions in space and that they can be shared with another atom.

The motion and energy of these bonding electrons will be determined not just by their parent atoms but also by:

> **(i) their interactions with the positive nuclei of the atoms they bond;**
>
> **(ii) their interactions with one another.**

The significance of the tendency of electrons and of other sub-atomic particles to pair was first noted by the chemist Lewis in 1916.

…whenever a radioactive atom emits one beta-particle [high-energy electron] it almost immediately emits another; again illustrating the instability of an unpaired electron within the nucleus. So also we find that in all the more stable states which atoms assume, the electrons occur in even numbers in the several inner shells ... the valence electrons almost invariably follow the same rule. The simplest explanation of these facts appears to lie in the assumption of a physical pairing of the electrons. There is nothing in the known laws of electric force, nor is there anything in the quantum theory of atomic structure, so far as it has yet been developed, to account for such a pairing ... there can be no question that the coupling of electrons is intimately connected with the magnetic properties of the electron orbits, and the explanation of this phenomenon must be regarded as one of the most important outstanding problems in quantum theory.

As we shall see, the explanation for this phenomenon, which is central to a fuller understanding of atomic and nuclear structure and of the nature of the chemical bond, was only found some fifteen years later, when quantum mechanics was applied to atomic theory.

Part 3.3 The Quantum Mechanical Model of the Atom.

Heisenberg had argued that in a consistent system of atomic mechanics, no entities should be introduced that are not physically observable.

In his judgement, the Bohr model broke down because in addition to dealing with the observable frequencies of the radiations emitted and absorbed by an atom, it also spoke of electron trajectories.

However, as we have seen, minute particles such as electrons are inherently elusive.

Fixed paths, orbits or trajectories cannot be assigned to them; their motion can only be construed in terms of probabilities.

A consistent atomic mechanics must therefore describe the motion of atomic electrons in terms of measurable probabilities and not in terms of trajectories.

In principle, these probabilities can be computed from the wave functions (probability amplitudes) of the atomic electrons which, in turn, can be obtained from the solutions of the appropriate Schrodinger equations.

The electrons are held in an atom by the electric field generated by the positive nuclear charge; they lack the necessary energy to escape this attraction.

Their situation is analogous to that of a particle confined to a small region of space -a three dimensional potential well.

As we have seen, a minute particle that is confined to a potential well can only occupy certain quantum states whose energies are sharply observable.

These stationary quantum states are characterized by integers called *quantum numbers*.

A three dimensional potential well allows for three integer quantum numbers.

In the case of an atomic electron, the first of these integers - designated n - fixes the particle's energy eigenvalues.

The other two quantum numbers, l and m_l, are associated with the electron's orbital angular momentum: l determines the magnitude of the orbital angular momentum, m_l its component parallel to an arbitrarily chosen axis.

Thus, each atomic stationary state is characterized by a set of three parameters, the value of each being that enjoined by its particular quantum number.

This is the essence of the quantum mechanical model of the atom.

Determining the quantum numbers that characterize the allowed states of the atomic electrons is the main task faced in developing the quantum mechanical model of the atom.

We will illustrate how this is done by first considering its application to the hydrogen atom.

3.3.1 The Hydrogen Atom

In developing a quantum mechanical model of hydrogen and hydrogen-like atoms, we begin, as we did with the Bohr model, by assuming that the interaction between the atom's negative electron and its positive nucleus (proton) is Coulombic in nature, i.e., that the attractive force, F_E, acting between the electron and the nucleus is given by the equation:

$$F_E = -\frac{kZe^2}{r^2} \tag{3.32}$$

where r is the distance between them and k is the electrostatic constant of proportionality.

The electron's potential energy, U, is given by:

$$U = -kZe^2/r \tag{3.33}$$

Substituting this value in the general time-independent Schrodinger equation for a particle of mass m_e gives

$$\frac{-h^2}{8\pi^2 m_e}\left(\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}\right)=\left(E+\frac{kZe^2}{r}\right)\psi \tag{3.34}$$

In view of the spherical symmetry of the potential field, it is more convenient to denote the position of the electron in the hydrogen atom in terms of the polar co-ordinates, r , θ , and ϕ , rather than the Cartesian co-ordinates, x, y and z (Fig 3.18).

The mathematical relationship between the two sets of coordinates is::

$$\begin{aligned} x &= r\sin\theta\cos\phi \\ y &= r\sin\theta\sin\phi \\ z &= r\cos\theta \end{aligned} \tag{3.35}$$

Placing the origin of the co-ordinate system at the hydrogen atom's nucleus, makes the co-ordinate r equal to the distance between the atom's positive nucleus and its electron.

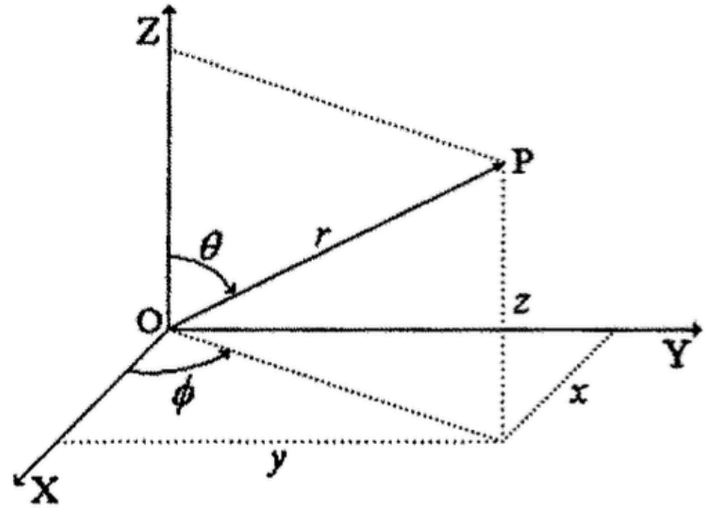


Fig 3.18 The Cartesian coordinates of the point P are (x, y, z). The polar coordinates are of the point are (r, θ, ϕ). θ is a 'latitude'; ϕ a 'longitude'.

The use of polar coordinates does not affect the total number of quantum numbers needed to describe the electron's allowed spatial states.

Irrespective of the co-ordinates used to describe its motion, three quantum numbers will be required because the electron moves in three dimensional space; there are three sets of boundary conditions that the electron's wave function must obey.

Likewise, the requirement that the wave function must have only one value at each point in space must still be satisfied.

Converting this equation into the corresponding polar co-ordinates equation by substituting the appropriate partial derivatives obtained from equations (3.35) gives:

$$\left[\frac{\partial^2\psi}{\partial r^2}+\frac{2}{r}\frac{\partial\psi}{\partial r}+\frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial\psi}{\partial\theta}\right)+\frac{1}{r^2\sin\theta}\left(\frac{\partial^2\psi}{\partial\phi^2}\right)\right] + \frac{2m_e}{\hbar^2}\left(E+\frac{kZe^2}{r}\right)\psi=0 \tag{3.36}$$

This awesome differential equation can be solved by assuming that the electron's wave function, $\psi(r, \theta, \phi)$, can be expressed as the product of three separate functions, R(r), $\Theta(\theta)$ and $\Phi(\phi)$, each of which depends explicitly on just a single variable:

$$\psi(r,\theta,\phi) = R(r)\cdot\Theta(\theta)\cdot\Phi(\phi) \tag{3.37}$$

* $R(r)$, called the *radial wavefunction*, depends just on r;
* $\Theta(\theta)$ depends only on the zenith angle, θ;
* $\Phi(\phi)$ depends on the azimuth angle, ϕ, alone.

However, instead of performing this laborious procedure, we will use our physical intuition to reveal some of the more important results of the quantum mechanical model of the hydrogen atom.

The electric force that binds the electron to the nucleus of the hydrogen atom acts along the line joining them; it is a central force.

In a classical two-body system governed by a central force, two dynamical quantities are conserved: the system's total energy, E, and its orbital angular momentum, L, about the force center.

We might thus expect that in a hydrogen atom, both of these quantities will be sharply observable and quantized.

As regards the electron's total energy, we have already seen that the expression given by the Bohr model for the energy of the stationary states in the hydrogen atom,

$$E_n = -\frac{k^2 Z^2 m_e e^4}{2n^2 \hbar^2} \quad (n = 1, 2, 3, \ldots)$$

correctly predicts the observed frequencies of its spectral lines

Moreover, it satisfies the Correspondence Principle; it makes the same predictions as classical physics when applied to those macroscopic systems for which classical physics is valid.

Thus, we might well expect that the quantum mechanical model should give the same expression.

And indeed, the solution of the radial wave function, R(r), requires that energy eigenvalues, E_n, be assigned to the bound electron that correspond exactly to those given by the Bohr model.

Thus, the first of the quantum numbers, n, has the same role in the quantum mechanical model of the hydrogen atom as the principal quantum number in the Bohr model, namely, determining the electron's total energy.

The atomic energy levels described by the principal quantum number are often designated by the spectroscopic notation K, L, M,... associated with X-ray spectra (Table 3. 1).

Table 3.1: The spectroscopic notation of the principle states in the hydrogen atom

Principal quantum number	Spectroscopic notation
$n = 1$	K
$n = 2$	L
$n = 3$	M
$n = 4$	N

Turning to the electron's orbital angular momentum, we find that obtaining eigenvalues for this classically conserved dynamical variable, without contravening the uncertainty principle, restricts the way the other two quantum numbers can be used.

For example, if the direction of the orbital angular momentum vector, **L**, was precisely known, it could be chosen to define the orientation of a coordinate axis, say the Z-axis (Fig 3.19).

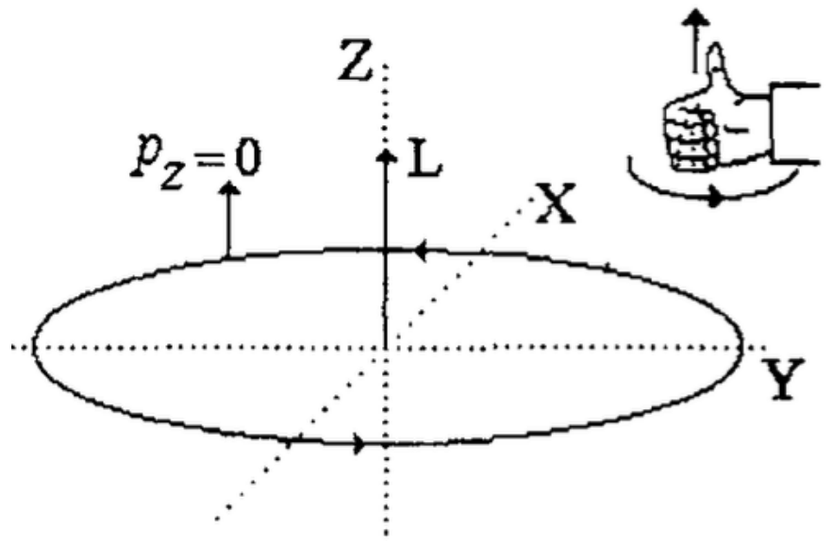


Fig 3.19 The orientation of the Z-axis is defined by that of the axial vector, **L.** This confines the electron to the orbital plane XY which in turn fixes the value of both its z co-ordinate and the z component of its linear momentum at zero.

This would confine the electron to orbits in the X Y plane and would unequivocally fix its z co-ordinate as equal to zero, implying a zero uncertainty, $\Delta z = 0$, in its z co-ordinate.

Moreover, under these circumstances, the electron could have no linear momentum directed out of the orbital plane.

This, in turn, would unequivocally set the z component, p_z, of the electron's linear momentum as equal to zero, implying a inherent zero uncertainty, $\Delta p_z = 0$, in its value.

Clearly, a situation where both uncertainties, Δz and Δp_z, equaI zero contravenes the uncertainty principle, $\Delta z \, \Delta p_z \geq \hbar$.

This is precisely the situation in the Bohr model of the hydrogen atom and illustrates its fundamental dissonance from quantum mechanics.

Evidently, precise knowledge of the direction of the orbital angular momentum vector, **L**, is forbidden by the uncertainty principle.

What then are the sharp observables associated with the electron's angular momentum that can be characterized by the two remaining quantum numbers?

What is the most information we can possess about the electron's angular momentum vector whilst still remaining ignorant as to its direction in space?

Clearly, we cannot know the value of all three components, L_x, L_y, and L_z, of the electron's angular momentum for that would be equivalent to knowing the direction in which the vector **L** is pointing.

However, irrespective of its direction in space, or of the coordinate system used to define it, the magnitude of the electron's angular momentum will always satisfy the relationship:

$$L^2 = L_x^2 + L_y^2 + L_z^2 \tag{3.38}$$

Thus, the first of the sharp observables associated with the electron's angular momentum is the square of its magnitude, L^2.

The full solution of the Schrodinger equation for the hydrogen atom shows that the eigenvalues of the square, L^2, of the magnitude of the orbital angular momentum of an atomic electron are:

$$L^2 = l(l+1)\hbar^2 \qquad \text{where } l = 0, 1, 2, 3\ldots, (n-1) \tag{3.39}$$

l is called the orbital quantum number. It follows, that in its ground state, n = 1, the orbital quantum number of the electron in atomic hydrogen is zero: $l = 0$.

Thus, its angular momentum in this state is also equal to zero.

This and not the value $\hbar$, given by the Bohr model, is the one borne out by experiment.

The atomic sub-levels characterized by the l quantum numbers are also often designated by the spectroscopic notation s, p, d, *f* and g (Table 3.2).

Table 3.2: The spectroscopic notation of the quantum states in the hydrogen atom

	$l = 0$	$l = 1$	$l = 2$	$l = 3$	$l = 4$
$n = 1$	1*s*				
$n = 2$	2*s*	2*p*			
$n = 3$	3*s*	3*p*	3*d*		
$n = 4$	4*s*	4*p*	4*d*	4*f*	
$n = 5$	5*s*	5*p*	5*d*	5*f*	5*g*
$n = 6$	6*s*	6*p*	6*d*	6*f*	6*g*

Ignoring relativistic effects, all states in the hydrogen atom designated by the same principal number, n, possess the same energy, whatever their orbital quantum number, l.

This congruence of energy levels is called *degeneracy*.

We say that sub-levels sharing the same energy are *degenerate*.

However, each degenerate state is described by a different wave function.

What information can the third quantum number give us?

Accurate knowledge (consistent repeated measurements) regarding of any one of the components of the angular momentum, say L_z, will still leave us ignorant as to the actual direction in which the vector **L** is pointing for we still have no information regarding the components L_x and L_y.

This is equivalent to saying that we have no reliable information about the azimuth angle, ϕ, the angle in the X Y plane.

In terms of the uncertainty principle this can be expressed as:

$$\Delta\phi \Delta L_z \approx \hbar \tag{3.40}$$

Thus, the second of the sharp observables associated with the electron's angular momentum is its component parallel to an arbitrarily chosen Z-axis.

From the solution to the Schrodinger equation, we obtain that the eigenvalues of the L_z component of the orbital angular momentum of an atomic electron are given by:

$$L_z = m_l \hbar$$
$$\text{where } m_l = -l, (-l+1), (-l+2) \ldots, l \tag{3.41}$$

m_l is called the magnetic quantum number.

It follows, that there are $(2l + 1)$ sub-states corresponding to each l state.

Thus, whereas the $l = 0$ state allows for only one sub-state, $m_l = 0$, the $l = 2$ state permits five sub-states, $m_l = -2, -1, 0, +1, +2$.

Table 3.3: The sub-states in the hydrogen atom

	m_l
$l = 0$	0
$l = 1$	−1,0,+1
$l = 2$	−2, −1,0,+1,+2
$l = 3$	−3,−2, −1,0,+1,+2,+3
$l = 4$	−4,−3,−2, −1,0,+1,+2,+3,+4

The picture that emerges from these two quantum numbers is of the angular momentum vector, **L**, precessing around the designated Z-axis at an angle θ given by:

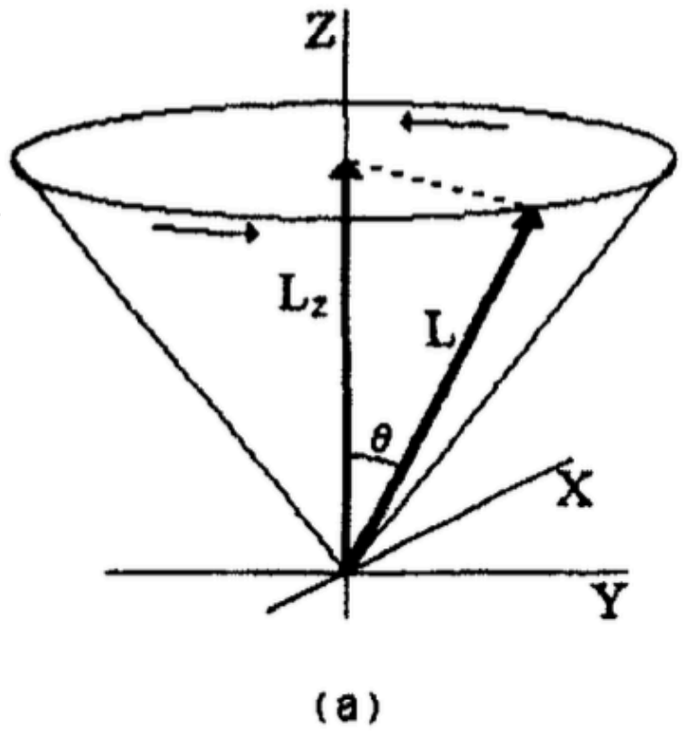
$$\cos\theta = \frac{L_z}{L} = \frac{m_l}{\sqrt{l(l+1)}} \tag{3.42}$$

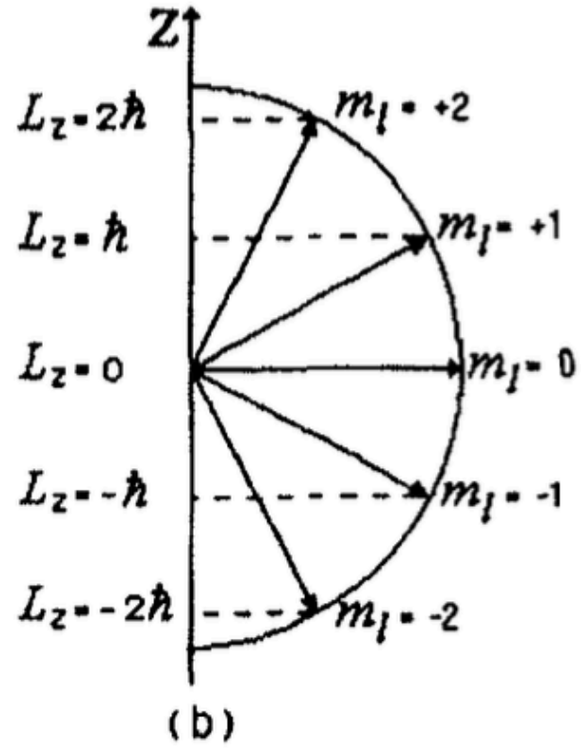
The angle θ is restricted to certain discrete values in accordance with the values of the quantum numbers l and m_l (Fig 3.20).

Fig 3.20 The vector model of angular momentum quantization. (a) The angular momentum vector, **L**, precesses around the Z-axis at certain discrete angles given by

$$\cos\theta = L_z/L$$

In this visualization, the components L_x and L_y are changing continuously. (b) The eigenvalues and orientations of the angular momentum vector in the case where $l = 2$, i.e., L = h and $L_z = m_l h$.





It must be emphasized, that this space quantization predicted by the quantum mechanical model of the atom, is not a physical restriction on the orientation of the orbital plane relative to the direction of the magnetic field; there is no actual precessional motion of the electron.

Space quantization is a fundamental property of space itself.

It makes no difference how the direction of the Z-axis is defined; whether by reference to an external magnetic field or to any other effect, space quantization will be observed.

We can now use the three quantum numbers, n, l and m_l, that we have devised in order to build a quantum mechanical model of the hydrogen atom.

In terms of its quantum numbers, the ground state of the hydrogen atom is characterized by the set: n = 1, $l = 0$, $m_l = 0$.

In order to indicate the quantum states it describes, an electron's wave function, $\psi(r,\theta,\phi)$,is usually designated by $\psi_{n,l,ml}$.

Thus, the ground state's wave function is designated by $\psi_{1,0,0}$.

The solution of the Schrodinger equation for the ground state of the hydrogen atom gives the wave function:

$$\psi_{1,0,0} = \frac{1}{\sqrt{\pi a_0^3}} e^{-r/a_0} \tag{3.43}$$

where a_0 is the Bohr radius.

Note that this function is independent of both the zenith and azimuth angles, i.e., it is spherically symmetrical.

Consequently, the probability of finding a ground state electron at a distance r from the nucleus will be the same in all directions.

In the ground state, the probability density at any point at a distance r from the hydrogen atom's nucleus is $|\psi_{1,0,0}|^2$.

Because of the spherical symmetry of the wave function, the overall probability density, P(r), at a distance r from the nucleus is equal to the probability density at any one point, $|\psi_{1,0,0}|^2$, multiplied by the surface area, $4\pi r^2$, of a sphere of radius r:

$$P(r) = |\psi_{1,0,0}|^2 4\pi r^2 \tag{3.44}$$

P(r) is called the *radial probability density*. Substituting the expression for the ground state wave function, $\psi_{1,0,0}$, gives:

$$P(r) = \left(\frac{1}{\sqrt{\pi a_0^3}} e^{-r/a_0}\right)^2 4\pi r^2 = \frac{4r^2}{a_0^3} e^{-2r/a_0} \tag{3.45}$$

This function has its maximum value when $r = a_0$.

Thus, in the ground state the probability of finding the electron in the hydrogen atom is greatest at a distance a_0 from the nucleus.

This is the significance of the Bohr radius in quantum mechanics (Fig 3.21).

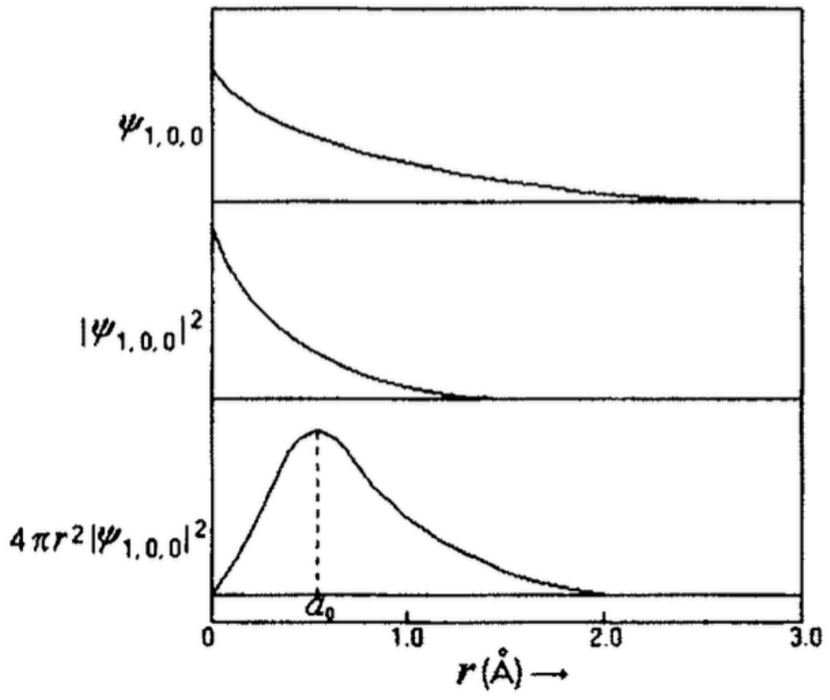


Fig 3.21 The wave function, $\psi_{1,0,0}$ and probability density, $|\psi_{1,0,0}|^2$ and radial probability density, $P(r) = |\psi_{1,0,0}|^2$, of the electron in a hydrogen atom in the ground state.

The radial probability density, P(r), in excited hydrogen atom states whose orbital quantum number $l = 0$, e.g., the 2s, 3s, ..., ns states, is also spherically symmetrical.

However, the radial probability density in these states does show distinct maxima and minima at different distances from the nucleus (Fig 3.22).

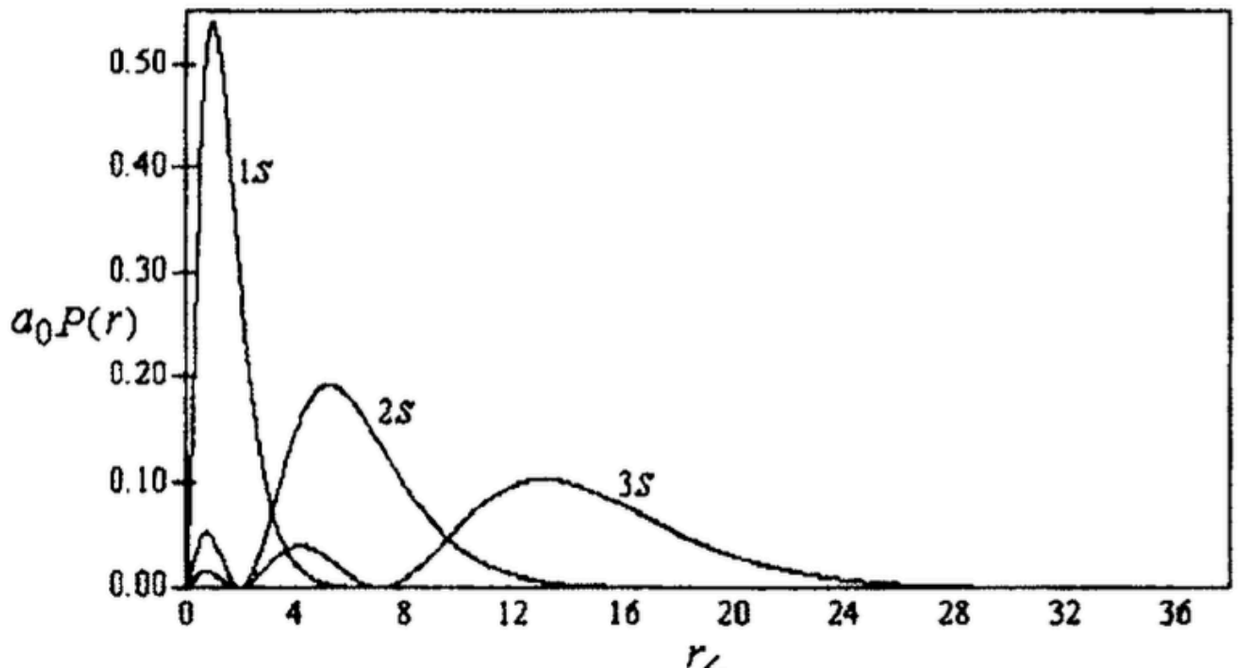


Fig 3.22 The radial probability density functions of the Is, 2s and 3s states in the hydrogen atom.

In general, the higher the principal quantum number the further away from the nucleus is the largest probability density maximum.

In states having $l > 0$ there is no spherical symmetry (though there may be other symmetries).

In these states, the probability density also depends on the zenith and azimuth angles.

The 2p state (n = 2, l = I), in which the magnetic quantum number can take the values m_l = -1, 0 or +1, exhibits distinct directional characteristics.

For example, in the $\psi_{2,2,0}$ state, the electron is found preferentially along the Z-axis; for this reason, this state is designated the $2p_z$ state.

The wave functions of the other two 2p states, $\psi_{2,2,+1}$ and $\psi_{2,2,-1}$, can combine by addition or by subtraction to give two more wave functions, $\psi_{2,px}$, and $\psi_{2,py}$, that possess distinct directional preferences:

$$\psi_{2p_x} = \frac{1}{\sqrt{2}}\left(\psi_{2,1,+1} + \psi_{2,1,-1}\right)$$
$$\psi_{2p_y} = \frac{1}{\sqrt{2}}\left(\psi_{2,1,+1} - \psi_{2,1,-1}\right) \qquad (3.46)$$

Assigning the same amplitude to each of the component waves, $\psi_{2,2,+1}$ and $\psi_{2,2,-1}$, indicates that each makes an equal contribution to the new wave functions.

The preferred motion of the electron in the $\psi_{2,px}$,function ($2p_x$ state) is along the X-axis, and that in the $\psi_{2,py}$ function ($2p_y$ state) is along the Y-axis (Fig 3.23).

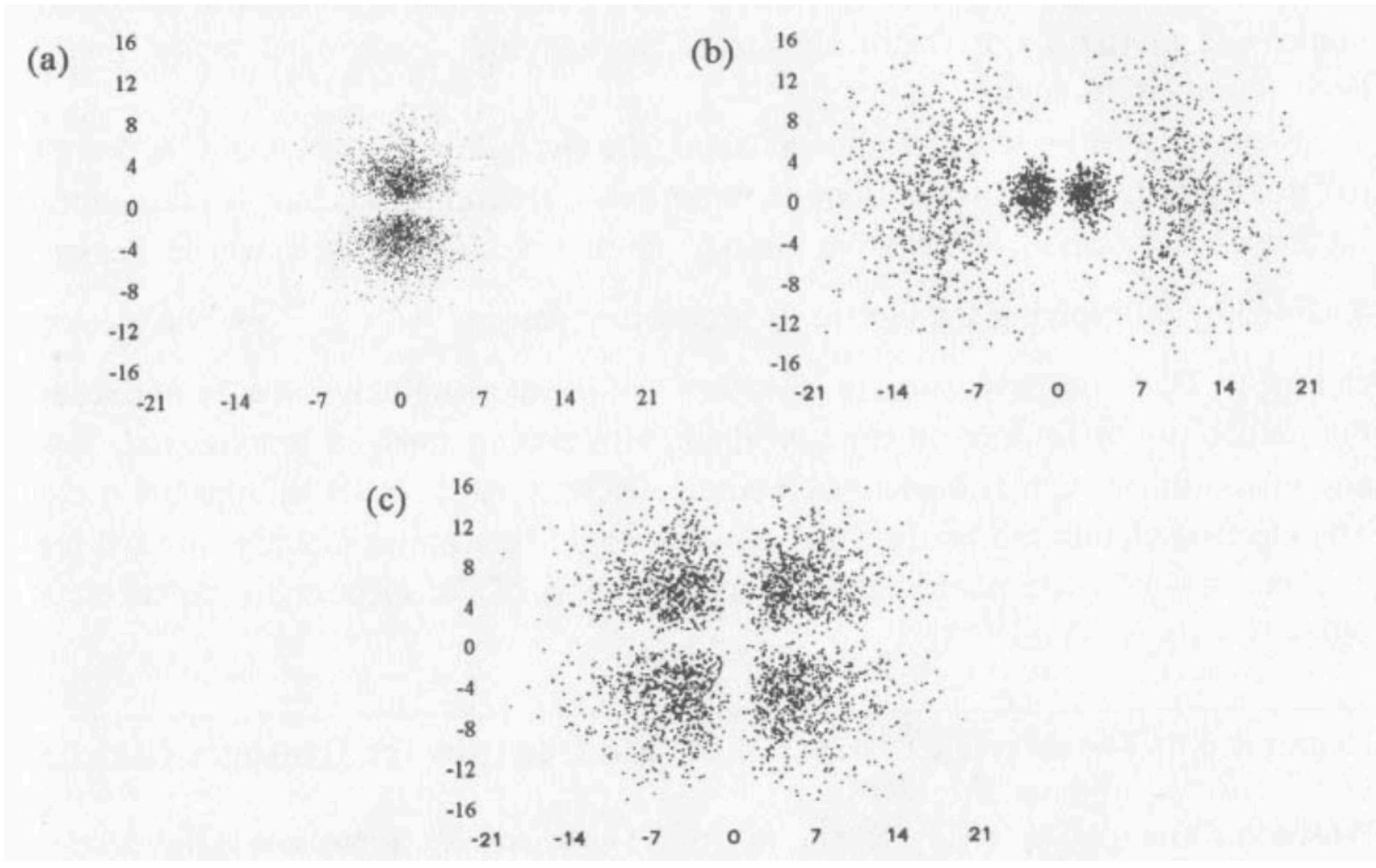


Fig 3.23 Computer simulations of the image that would be revealed on looking through two thousand slide photographs (transparencies) of the electron in a hydrogen atom in (a) the $2p_0$ state ($\psi_{2,1,0}$); (b) the $3p_{\pm1}$ state ($\psi_{2,1,\pm1}$); (c) the $3d_{\pm1}$ state ($\psi_{3,3,\pm1}$) The images were produced by plotting the probability density, $|\psi_{n,l,m_l}|^2$ of each state at two thousand randomly selected points in the zx plane. The scaling shows the distance from the nucleus in Bohr radii.

This directional preference is of great import in the explanation of chemical bonding as it affords the possibility of directed bonds and of molecular architecture.

As such, it marks one of the major improvements of the quantum mechanical model over the Bohr-Sommerfeld model.

Since, according to quantum mechanics, the electron's motion cannot be described in terms of trajectories, its distribution in the space around the nucleus is often pictured in terms of a 'cloud of negative charge' or an 'electron cloud', whose density is, obtained by multiplying the electron's probability density, $|\psi_{n,l,m_l}|^2$ by the electronic charge, e.

The ‘charge density’ or ‘thickness’ of the electron cloud at a point expresses the probability of finding an electron there.

However, it must be emphasized, that in any measurement that is made, just a single electron will be revealed and not a cloud.

The electron clouds can be regarded as what would be seen on looking through a pile of thousands of slide photographs (transparencies) of the electron, each taken at a different moment (Fig 3.23).

3.3.2 Atomic Spectra and Quantum Mechanics

Transitions between stationary states in which radiation is emitted or absorbed are called *radiative transitions*; they are the transitions that appear as spectral lines.

The Bohr model relates the frequency of the spectral lines to the difference between the energies of the stationary states.

However, it offers no real explanation for the actual occurrence of radiative transitions.

Why should atoms spontaneously shift from one stationary state to another?

Apart from the intuitive feeling that systems ought to ‘fall’ from higher energy states to lower ones, why should excited atoms spontaneously ‘drop’ from higher stationary states into lower ones.

On the contrary, it can be argued that since an atom's stationary states are all non-radiating, they are all equally 'desirable'.

Moreover, other orbiting bodies such as planets don't spontaneously change their orbits, so why should orbiting electrons?

And if electrons do in fact spontaneously 'drop' from higher stationary states into lower ones, how are the photons emitted in the transfers actually created?

The spontaneous emission of radiation has no classical analogue.

It is a purely quantum phenomenon, a satisfactory explanation for which can only be given in terms of the quantum mechanical model of the atom.

It is most readily understood as arising from the superposition of atomic wave functions.

The electron involved in a radiative transition can occupy either of the quantum states, i and j, between which it moves.

There is no way we can tell which state it is occupying at any moment, the higher or the lower one.

Thus, the electron's time dependent wave function, $|\psi_{ij}|^2$, is given by the linear combination (superposition) of the time dependent wave functions, ψ_i , and ψ_j, corresponding to the two states:

$$\Psi_{ij} = A_i \Psi_i + A_j \Psi_j \qquad (3.47)$$

where A^2_i and A^2_j are the probabilities of the electron being found in each state.

Normalization requires that $A^2_i + A^2_j = 1$ at all times.

When the electron is in the i state, $A_i = 1$ and $A_j = 0$; when it is in the j state, $A_i = 0$ and $A_j = 1$.

In each of these instances, which correspond, respectively, to the i and j stationary states, the probability density, $|\Psi_{ij}|^2$, is constant.

But what happens to the probability density when both A_i and A_j have non-vanishing values?

In general, the time dependent wave function, $\Psi_s(x, y, z, t)$, of a particle in a stationary state can be expressed in terms of its time independent wave function , $\psi(x, y, z)$, and its energy eigenvalue, E, in the form

$$\Psi_S(x, y, z, t) = \psi(x, y, z)e^{-iEt/\hbar} = \psi(x, y, z)e^{-i2\pi ft} \tag{3.48}$$

where, $f = E/h$ is the frequency of the particle's matter wave. This gives for the wave functions of the i and j stationary states:

$$\Psi_i = \psi_i e^{-i2\pi f_i t} \text{ and } \Psi_j = \psi_j e^{-i2\pi f_j t} \tag{3.49}$$

where $f_1 = E_1/h$ and $f_2 = E_2/h$ are respectively, the frequencies of the matter wave of the i state and the j state. Substituting these expressions in equation (3.47) gives

$$\Psi_{ij} = A_i\psi_i e^{-i2\pi f_i t} + A_j\psi_j e^{-i2\pi f_j t} \tag{3.50}$$

Compared to the frequencies of the matter waves, the difference between them f_i - f_j, is very small.

Consequently, when both A_i and A_j have non-vanishing values, the superposition of the matter waves corresponds to the phenomenon of beats, the periodic amplitude modulation that arises from the superposition of two vibrations with neighboring frequencies (Fig 3.24).

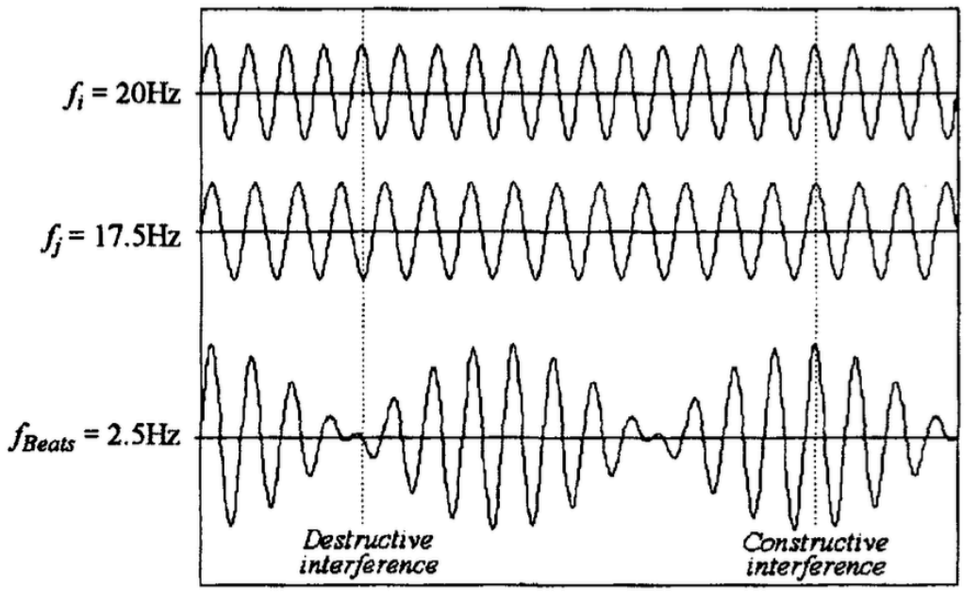


Fig 3.24 The beats produced by the superposition, $\Psi_{ij}=\Psi_i+\Psi_j$, of two waves with similar frequencies: $f_i = 20Hz$; $f_j = 17.5$Hz. Hence, $f_{beats} = 2.5Hz$

The frequency, f, of the beats is equal to the absolute value of the difference in the frequencies of the two component waves.

In this case:

$$f = |f_i - f_j| \tag{3.51}$$

Since, in general, the probability density of a state depends on the amplitude of its wave function, the probability density of the ij state, $|\Psi_{ij}|^2$,which is usually called the *transition density*, will oscillate with the same frequency as does the amplitude of the resultant wave, i.e., with the beat frequency , $f = |f_i - f_j|$

In a stationary state, the charge density, e $|\Psi_{ij}|^2$, associated with the electron does not oscillate and so radiation is neither emitted nor absorbed.

However, when the electron is in transit between the two states, i.e., when both A_i and A_j, have non-vanishing values, the charge density oscillates with the beat frequency.

These fluctuations in the electronic cloud around the nucleus, that arise from the superposition of the matter waves of the stationary states, are the quantum equivalent of the electric oscillations that, according to classical electromagnetic theory, must accompany the emission and absorption of radiation.

Only when these fluctuations arise can radiation be emitted or absorbed.

Thus, the frequency, f, of the radiation emitted or absorbed in a transition is determined by that of the periodic amplitude modulation (the beats) that arises from the superposition of the matter waves of the stationary states between which the transition occurs.

$$f = f_i - f_j = \frac{E_i - E_j}{h} \tag{3.52}$$

In this way, Bohr's equation, $f_{ij} = \frac{E_i - E_j}{h}$ that was introduced as a postulate in Bohr's model of the atom, arises as a straightforward corollary of the wave nature of the electron in the quantum mechanical model of the atom.

However, the generation of a fluctuating charge density -an oscillating electric dipole- is not the only condition that must be satisfied for the emission or absorption of electromagnetic radiation, i.e., of photons.

It transpires that radiative transitions never occur when the total angular momentum of both the initial and final states of the emitting or absorbing system equals zero.

Assuming that angular momentum is conserved in these transitions, this indicates that the photons emitted or absorbed in radiative transitions must themselves possess angular momentum.

The classical analogue of this intrinsic angular momentum of the photons is the circular polarization of the electromagnetic wave.

It follows, that radiative transitions can only occur when the fluctuations in the electron cloud and the attendant emission or absorption of radiation obey the law of angular momentum conservation.

Intense spectral lines correspond to radiative transitions whose probability is high; faint lines correspond to transitions of low probability; forbidden transitions are those whose transition density equals zero.

Allowed atomic transitions in hydrogen and hydrogen-like atoms obey the selection rules:

$$\Delta l = \pm 1 \qquad \Delta m_l = 0, \pm 1$$

Both rules ensure the conservation of angular momentum. In addition, the $\Delta l = \pm 1$ rule ensures that a non-vanishing electric dipole is generated by the fluctuations in the electron cloud (Fig 3.25).

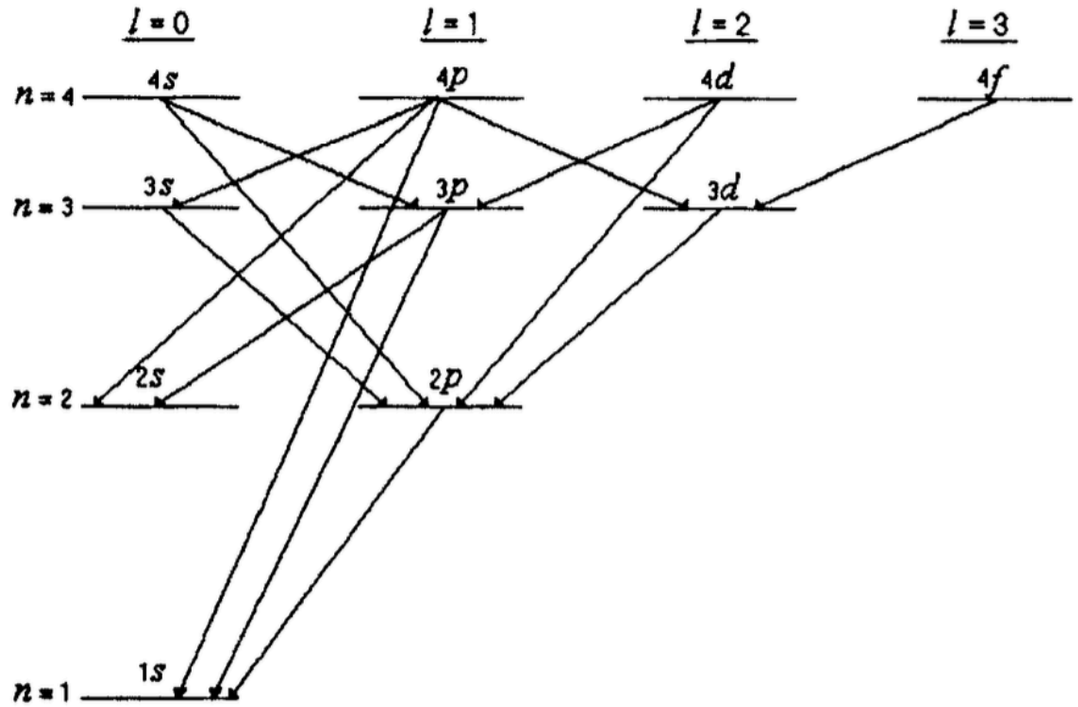


Fig 3.25 Examples of allowed radiative transitions in the hydrogen atom that obey the rule $\Delta l = \pm 1$.

Both rules ensure the conservation of angular momentum.

In addition, the $\Delta l = \pm 1$ rule ensures that a non-vanishing electric dipole is generated by the fluctuations in the electron cloud (Fig 3.25).

Similar selection rules had been previously derived empirically by investigators of the Bohr-Sommerfeld model.

They were explained at the time by applying the correspondence principle to classical harmonics; if the rules are valid for high quantum numbers, their validity can be postulated for all quantum numbers.

However, it was the quantum mechanical model that first gave them an a priori theoretical basis.

The average lifetime of an excited state is inversely proportional to the transition probability for a drop to a lower energy state.

The average lifetime of a typical excited atomic state is ~10^{-8}s. The lifetime of the ground state is practically infinite for there is no energy level below it.

Of the excited states in the hydrogen atom, the 2s state is of particular interest.

Being an excited state, its lifetime might be expected to be very short.

However, there is no lower state to which the selection rule $\Delta l = \pm 1$ allows it to fall; the only state below it is the 1s state which shares the same I = 0 quantum number.

Consequently, the 2s state has a longer than expected lifetime and is thus called a *metastable* state.

To summarize the quantum mechanical model of the hydrogen and hydrogen-like atoms:

1. Hydrogen and hydrogen-like atoms consist of a very small and very dense nucleus that contains the entire positive charge of the atom, $+Ze$, and virtually all of its mass.
2. A single negative electron, to which no specific trajectory can be assigned, occupies the space around the nucleus: this electron, together with the nucleus, comprises the atom.
3. The interaction between the nucleus and the electron is Coulombic.
4. The atom is described by a wave function which reflects the restrictions imposed by the Heisenberg uncertainty principle and from which the quantum numbers that characterise its stationary states can be derived.
5. The atom absorbs and emits energy in accordance with the probabilities prescribed by its wave function.

Extending the quantum mechanical model to the fine details of the spectra and to multi-electron atoms requires the introduction of a new concept - electron spin.

Part 3.4 Electron Spin

In 1925, two graduate students, Goudsmit and Ulenbeck, suggested that apart from any orbital angular momentum they might have, electrons also possess an intrinsic (internal) angular momentum - *spin*.

They were apparently unaware that this singular hypothesis, which provided a physical interpretation for the fourth quantum number that Pauli had already shown was needed to explain the observed splitting of fine structure spectral lines, had been suggested previously in 1921 by Richardson in order to account for the magnitude of the Amperian currents in magnetic materials.

Subsequently, the concept of electron spin became the basis for the elucidation of diverse phenomena such as the pattern of the periodic table of the elements and the pairing of electrons in chemical bonds.

3.4.1 Electron Spin

The intrinsic (internal) angular momentum of a spinning particle, its spin, is quantized in a way that is analogous to the quantization of orbital angular momentum.

The eigenvalues of the square, S^2, of the magnitude of a particle's spin are given by:

$$S^2 = s(s+1)\hbar^2 \tag{3.53}$$

where the quantum number s is called the *spin quantum number*.

The component, S_z, of the particle's spin parallel to an arbitrarily chosen Z-axis is given by:

$$S_Z = m_s\hbar$$
$$\text{where } m_s = -s,(-s+1),(-s+2),\ldots,s \tag{3.54}$$

m_s is called the spin magnetic quantum number.

It follows that the particle's intrinsic angular momentum vector can have (2s+ 1) orientations in space.

In the case of the electron, the quantum number s can take only one value: s = 1/2.

This allows for just two orientations of the spin vector corresponding to $m_s = \pm 1/2$: 'spin up'(↑) and 'spin down'(↓).

It is customary to call the quantum number s the *spin of the electron* and to say that "an electron has a spin of a half", this despite the fact that the magnitude of its spin is actually $S = \sqrt{3}\hbar/2$ (Fig 3.26).

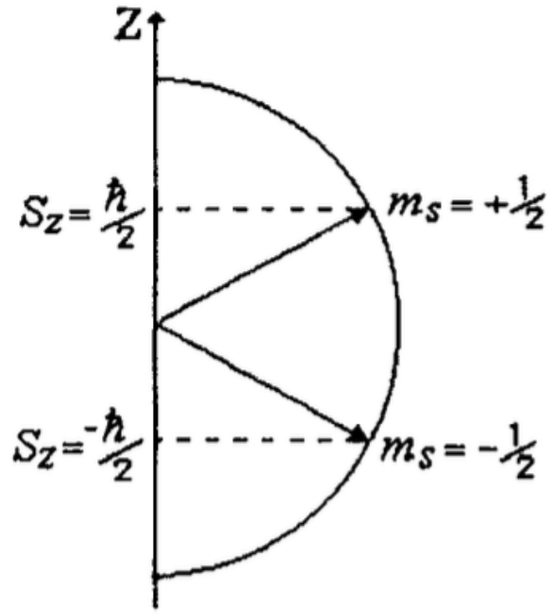


Fig 3.26 The space quantization of the electron's spin. In each orientation the magnitude of the spin is $S = \sqrt{3}\hbar/2$

It should be emphasized that it is incorrect to picture the electron as a spinning negatively charged solid sphere.

The electron's intrinsic angular momentum is not the result of the spinning of an inertial mass.

Electrons have no discernible dimensions; it is quite possible that they are point particles, Electron spin is a pure quantum phenomenon that is described by the *spin wavefunction*, χ_{ms}.

In 1929, Dirac showed that the existence of electron spin follows from the solution of the Schrodinger equation if the relativistic expression for the particle's energy, $E = \pm\sqrt{m_e^2c^4 + p^2c^2} + U$, is employed in its formulation instead of the classical expression, $E = p^2/2m_e + U$.

The picture this solution gives is of a flow of energy and momentum around the axis along which the electron's matter wave propagates.

It is this circulating flux that confers the intrinsic angular momentum on the electron.

In general, charged particles that possess an intrinsic angular momentum (spin), **S**, have a magnetic moment, μ_S, associated with them that is given by

$$\mu_S = \gamma_S \mathbf{S}\left(\frac{e}{2m}\right) \tag{3.55}$$

where m is the particle's mass.

The factor γ_s is called the *spin gyromagnetic ratio*; its value varies from particle to particle. In the case of an electron, $\gamma_s = -2.0024$.

The component, $\mu_{SZ(electron)}$, parallel to an arbitrarily chosen Z-axis of the magnetic moment associated with the electron's spin is given by

$$\mu_{S_Z(electron)} = \gamma_S S_Z\left(\frac{e}{2m_e}\right) \tag{3.56}$$

The space quantization of the electron's spin, up or down, results in a corresponding restriction on the orientations of the magnetic moment associated with its spin; it too is restricted to just two possible orientations.

This simple fact has profound repercussions.

3.4.2 The Stern-Gerlach Experiment

The phenomenon of space quantization was first predicted by the Bohr-Sommerfeld model of the atom, in which it arises from the quantization of the orbital angular momentum of the electron.

Its actual existence was verified experimentally by Stern and Gerlach in 1921, several years before the development of quantum mechanics.

However, as we shall see, what they actually observed was the space quantization associated with the electron's intrinsic angular momentum (spin) and not that which is a consequence of the quantization of its orbital angular momentum.

In a uniform field, the magnitude of the resultant force that acts on a dipole is zero.

Thus, though the dipole may rotate as a result of the couple impressed upon it by the field, particles possessing a dipole moment that enter a uniform field will pass through in a straight line (Fig 3.27a).

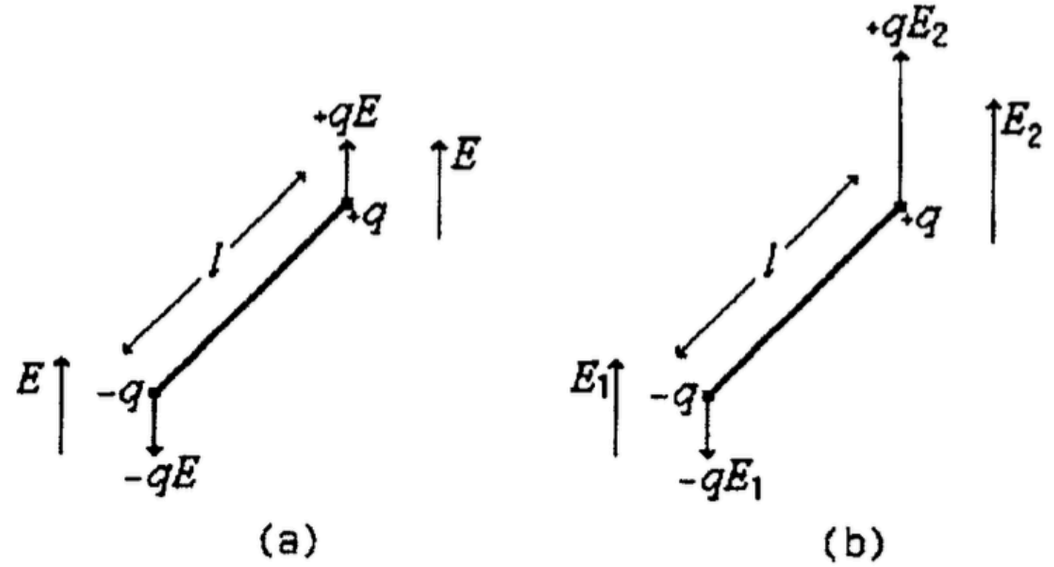


Fig 3.27 The forces acting on an electric dipole in an electric field. (a) In a uniform field the resultant force is zero: $+qE+(qE) = 0$. (b) In a non-uniform field the resultant is not zero: $+qE_2 + (-qE_1) \neq 0$.

This is equally true for an electric dipole in an electric field or a magnetic dipole in a magnetic field, where we do not distinguish ‘point’ magnets or ‘magnetic charges’.

However, in a non-uniform field the magnitude of the resultant force acting on the dipole does not equal zero and so the dipole’s center of mass will be accelerated (Fig 3.27b).

Thus, particles possessing a dipole moment that enter a non-uniform field will be deflected from their paths.

The extent of each particle’s deflection will depend on (i) the degree of non-uniformity of the field; (ii) the strength of the particle’s magnetic moment; (iii) the orientation of the particle’s magnetic moment relative to the direction of the field.

Many atoms behave like magnetic dipoles, i.e., as though they possess a magnetic moment.

According to classical theory, in the gaseous state these magnetic moments point randomly in all directions.

Consequently, passing a narrow collimated beam of gaseous neutral atoms through a non-uniform magnetic field should result in an overall broadening of the beam.

The atoms should be deflected in random directions, spreading the beam out evenly over a larger area.

However, space quantization implies that the orientations of the atomic magnets are restricted to a discrete number of directions relative to any chosen axis.

Depending on whether the Bohr-Sommerfeld or the quantum mechanical theory is applied, the magnetic moment of an atom possessing orbital quantum numbers k or l is limited, respectively, to (2k + 1) or (2 l + 1) orientations relative to any defined direction.

The most convenient reference axis is that defined by the direction of the magnetic field; this is usually designated the Z-axis.

Consequently, when passing through a non-uniform magnetic field, a narrow collimated beam of atoms possessing magnetic moments should split into an odd number of beams.

If the orbital quantum number, k or l, equals 1, the original beam should split into 3 beams; if it equals 2, there should be 5 beams, and so on.

If the orbital quantum number equals zero, there should be no beam splitting at all.

Stern and Gerlach produced a non uniform magnetic field by means of an electromagnet whose poles were differently shaped; one pole was flat and the other wedged.

A collimated beam of neutral silver atoms was passed through the field between the poles (Fig 3.28)

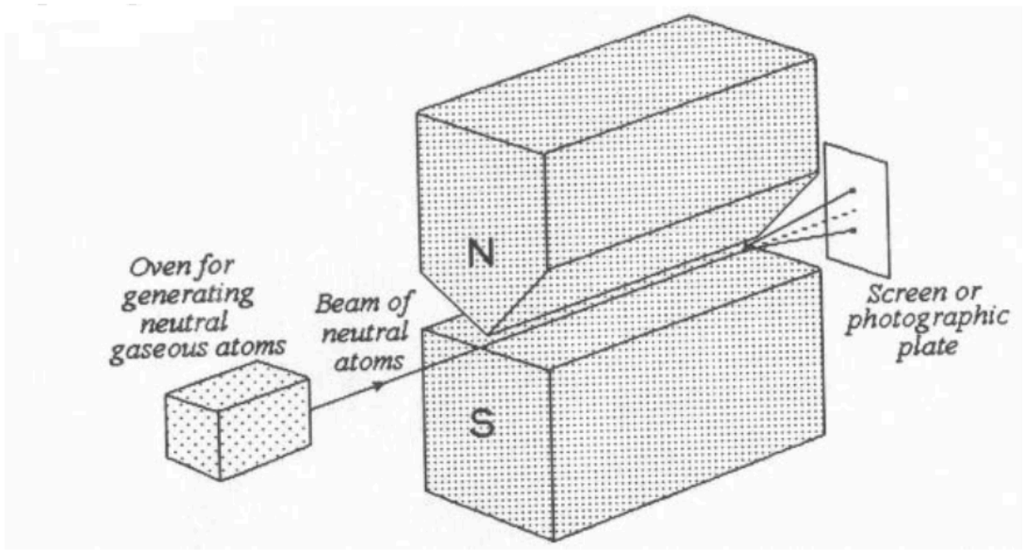


Fig 3.28 The Stern-Gerlach experiment. The collimated beam of neutral silver atoms passes through the non-uniform magnetic field and splits into just two beams.

They found that the beam split into two distinct beams.

Though this result demonstrated the quantization of space, the even number of beams, two in this case, was inexplicable at the time.

The experiment was repeated some years later with hydrogen atoms; they too divided into two beams.

In the Bohr-Sommerfeld model, the ground state in hydrogen has an orbital quantum number k = 1 and so the atoms should divide into three distinct beams on passing through the non-uniform magnetic field.

At first sight, the quantum mechanical model appears to fare no better.

It assigns an orbital quantum number $l = 0$ to the hydrogen atom's ground state.

This implies that the original beam should not split at all.

The splitting into two beams could be understood if, apart from its electron's orbital angular momentum, there was some other source of angular momentum in the hydrogen atom whose quantization was characterized by a fractional quantum number, 1/2.

Were this so, the number of beams produced would be 2x1/2 + 1 = 2, as found in the experiment, even when $l = 0$.

Fractional quantum numbers were completely beyond the conception of the Bohr-Sommerfeld model.

They are however anticipated by the relativistic solution of the electron's Schrodinger equation and by the matrix derivation of Heisenberg, both of which both assign the fractional quantum number, s, to the electron's intrinsic angular momentum (spin).

Thus, the splitting into two beams observed in the Stern-Gerlach experiment can only be explained by quantum mechanics, in which it is attributed to the space quantization associated with the spin of the electron.

3.4.3 Spin-Orbit Coupling

The two magnetic moments that an atomic electron can generate, that associated with its orbital motion and that associated with its spin, can interact.

In order to understand this interaction, we must remember that physics permits no preferred frame of reference; all reference frames are equivalent.

Thus, although we have tended to describe the atom in terms of a stationary nucleus and a moving electron, i.e., from the typical reference frame of the nucleus, an observer on the electron would be just as entitled to describe it in terms of the apparent motion of the nucleus relative to him.

Such an observer would note that, as a consequence of its motion around the electron, the charged nucleus induces a magnetic field in the region.

The interaction of this magnetic field with the spin magnetic moment of the electron affects the electron’s energy levels, causing some of them to split.

The phenomenon is a kind of internal Zeeman effect, the magnetic field with which the atomic electron interacts being one that is created by the atom itself.

Due to the coupling of the two magnetic moments, it is the electron's total angular momentum, $\mathbf{J} = \mathbf{L} + \mathbf{S}$, that is the physical quantity whose quantum numbers, j and m_j, determine the energy eigenvalues of the electron's quantum states.

The permitted values of the square, J^2, of the magnitude of the electron's total angular momentum is characterized by the values:

$$J^2 = j(j+1)\hbar^2 \qquad \text{where} \quad j = l \pm \tfrac{1}{2} \tag{3.57}$$

The component, J_z, of the electron's total angular momentum is given by:

$$J_z = m_j \hbar \qquad \text{where } m_j = -j, (-j+1), (-j+2) \ldots, j \tag{3.58}$$

The total angular momentum quantum number, j , is always a fractional number.

Consequently, the number, 2j + 1, of spatial orientations of the electron's total angular momentum vector, **J** , is always even.

For example, when $l = 1$, the total angular momentum quantum number can have two values: j = 1/2 or j = 3/2 giving two and four orientations, respectively.

The eminent example of spectral line splitting associated with spin-orbit coupling is the sodium doublet.

The yellow line in the sodium emission spectrum that corresponds to a transition from the 3p level to the 3s level is split into two.

The twin adjacent lines are attributed to a splitting of the 3p level into two sub-levels.

The upper 3p state corresponds to a total angular quantum number j = 3/2 ; the lower state to j = 1/2 (Fig 3.29).

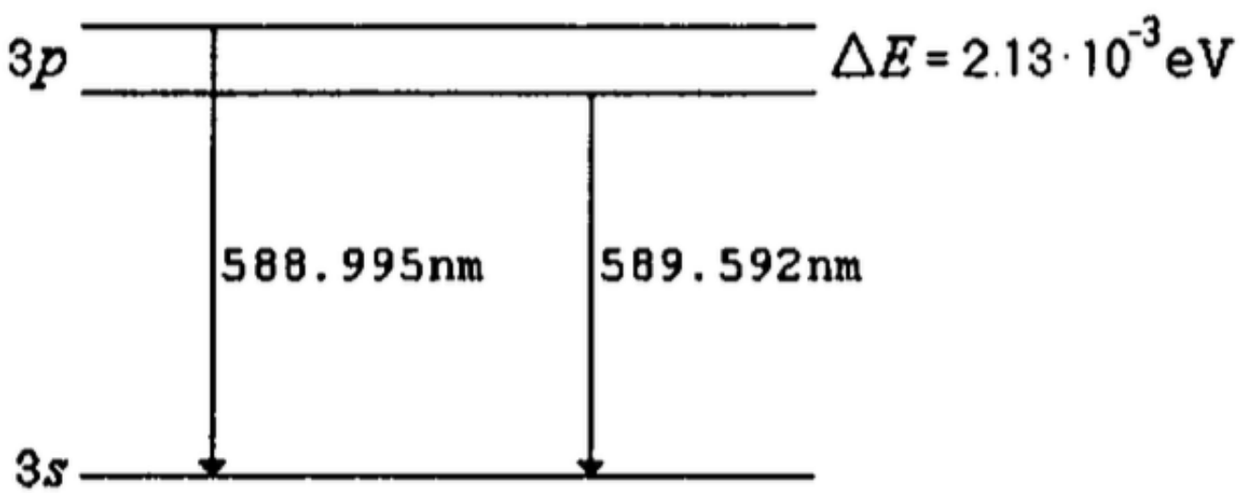


Fig 3.29 The sodium doublet. Two adjacent yellow spectral lines are observed corresponding to transitions from the upper and lower sub-levels of the 3p state to the 3s state. There is no splitting of the s level since no internal magnetic field arises in a state with zero orbital angular momentum. When $l = 0$ the only possible value of j is j = 1/2. Because the energy difference, ΔE, between the 3p sublevels is about one thousandth that between the 3p and 3s levels the doublet is only revealed at high resolutions.

A similar splitting is observed in transitions from the 2p state to the 1s state in hydrogen.

It follows, that the quantum numbers l (magnitude of the orbital angular momentum), s (magnitude of the spin) and j and m_j (magnitude and z component of the total angular momentum) are 'good' quantum numbers for the hydrogen atom.

Their values remain fixed so long as the electron remains in the same energy eigenstate.

Thus, repeated measurements of the quantities they define, which correspond to the constants of motion in classical mechanics, yield the same result every time.

On the other hand, the quantum numbers m_l and m_s, that define the z components of the orbital and intrinsic angular momentum respectively, are not good quantum numbers; their values change.

Repeated measurements of the z components they define yield one or other of the values $m_l = 0, f1, f2, \ldots$ or $m_s = f$, but which value emerges on each occasion is uncertain.

3.4.4 The Pauli Exclusion Principle and the Periodic Table

Solving the Schrodinger equation for an electron in a multi-electron atom presents great difficulties.

The nucleus of a neutral multi-electron atom of atomic number Z carries a charge of +Ze and is surrounded by a cloud of Z electrons.

Each electron is influenced by the electric fields produced by the positive nuclear charge and the other (Z - 1) negative electrons in the atom.

The potential energy function, $U(r, \theta, \phi)$, that describes such a situation is so complex that a complete solution of the Schrodinger equation for such an atom is unattainable.

The best that can be obtained is a series of estimates and approximations.

In making these estimates, it is reasonable to assume that in the ground state, each electron in the atom will occupy the lowest possible stationary state that it can.

This begs the immediate question of why don't all the electrons in an atom pile into the lowest quantum state, the 1s state, where they would all share the set of quantum numbers n = 1, $l = 0$, $m_l = 0$ and $m_s = \pm 1/2$.

That this is not the case is evident from the actual chemical and physical properties of the elements.

The chemistry of each of the three elements fluorine, neon and sodium is completely different even though they have almost the same number of electrons: 9, 10 and 11 electrons, respectively.

Were all their electrons in the same 1s quantum state, the addition of one more or the subtraction of one would hardly explain the obvious differences in their chemical properties.

On the other hand, the elements helium, neon, argon, krypton, xenon and radon, that have 2, 10, 18, 36, 54 and 86 electrons respectively, show striking similarities in all their properties, despite the considerable differences in the number of their electrons.

Clearly, the deployment of electrons in the atom involves more than just apportioning them all to the lowest quantum state.

The solution to this problem was found in the framework of the following principle *-the exclusion principle-* formulated by Pauli in 1925 as a result of his researches in atomic spectroscopy:

> **In a system containing a number of electrons, no two electrons can occupy the same state; each electron must have a different set of quantum numbers.**

In a system containing a number of electrons, no two electrons can occupy the same state; each electron must have a different set of quantum numbers.

Thus, allotting all the electrons in a multi-electron atom to the 1s state or to any other single state is forbidden.

The electrons must be arranged in the atom such that no two share the same set of quantum numbers; each electron in the atom must have its own unique set of the four quantum numbers, n, l, m_l and m_s.

Table 3.4 lists all the different atomic states (sub-levels) that can be specified by these four quantum numbers for values of the principle quantum number from n = 1 to n = 5.

Table 3.4 The electronic quantum states in an atom.

n	l	Spectroscopic symbol	m_l	Number of states[58]
1	0	1*s*	0	2
2	0	2*s*	0	2
	1	2*p*	−1,0,+1	6
3	0	3*s*	0	2
	1	3*p*	−1,0,+1	6
	2	3*d*	−2,−1,0,+1,+2	10
4	0	4*s*	0	2
	1	4*p*	−1,0,+1	6
	2	4*d*	−2,−1,0,+1,+2	10
	3	4*f*	−3,−2,−1,0,+1,+2,+3	14
5	0	5*s*	0	2
	1	5*p*	−1,0,+1	6
	2	5*d*	−2,−1,0,+1,+2	10
	3	5*f*	−3,−2,−1,0,+1,+2,+3	14
	4	5*g*	−4,−3,−2,−1,0,+1,+2,+3,+4	18

In the ground state of hydrogen (Z = 1), H, the first element in the periodic table Table 3.5 - the sole electron occupies the 1s state; n = 1, $l = 0$, $m_l = 0$. Its spin may be up or down.

Table 3.5: The periodic table of elements.

Group	I	II										III	IV	V	VI	VII	VIII
Period 1	1 H																2 He
2	3 Li	4 Be										5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg										13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	57-71	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	89-103	104 Rf	105 Ha	106 Sg	107 Bh	108 Hs	109 Mt									

Lanthanides	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
Actinides	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

The ordering of the electrons in the ground state of multielectron atoms, starts with the lowest energy state, the 1s state, and moves progressively from there up the energy scale until each of the atom's electrons has been assigned a unique set of quantum numbers.

Thus, in the ground state, both electrons in the inert gas helium (Z = 2), He, occupy the 1s state, but as required by the exclusion principle, each has a different spin quantum number; their spins are 'paired'. The respective quantum numbers of the helium electrons are:

	n	l	m_l	m_s
electron 1	1	0	0	$+\frac{1}{2}$
electron 2	1	0	0	$-\frac{1}{2}$

The electronic configuration of helium is designated by $1s^2$, i.e., two electrons in the 1s level.

No more than two electrons can occupy the 1s state; a larger number is precluded by the exclusion principle.

Thus, the third electron in lithium (Z = 3), Li, must be assigned to the level above, to the 2s state.

The electronic configuration of lithium is $1s^2 2s^1$, i.e., two electrons in the 1s state and one in the 2s state.

The two electrons in the 1s state constitute a complete or closed shell; this is the innermost K (n = 1) shell.

The single 2s electron is shielded from the nuclear charge by the closed K shell.

Consequently, the effective nuclear charge binding it to the atom is much less than +3e and its ionization energy is correspondingly small, just 5.4eV.

In general, an atom's chemistry is determined by the configuration of its outermost electrons, i.e., by those outside closed shells.

Lithium's chemical valency of +1 ensues from its single loosely bound 2s electron.

The electronic configuration of the four electrons in beryllium (Z = 4), Be, is $1s^2 2s^2$.

Because of the exclusion principle, the extra fifth electron in boron (Z = 5), B, cannot be added to the 2s sub-shell.

This electron must be placed in the 2p (l= 1, m_l = -l,0,l) sub-shell; this sub-level can accommodate a total of six electrons.

As a rule, the three m_l states of a p sub-shell are progressively filled in such a way as leaves as many unpaired electrons as possible; this is known as *Hund's rule*.

Table 3.6: Hund's rule.

Atom	Number of electrons	Electronic configuration	Spin of the p electrons
B	5	$1s^2\ 2s^2\ 2p^1$	↓
C	6	$1s^2\ 2s^2\ 2p^2$	↓ ↓
N	7	$1s^2\ 2s^2\ 2p^3$	↓ ↓ ↓
O	8	$1s^2\ 2s^2\ 2p^4$	↓↑ ↓ ↓
F	9	$1s^2\ 2s^2\ 2p^5$	↓↑ ↓↑ ↓
Ne	10	$1s^2\ 2s^2\ 2p^6$	↓↑ ↓↑ ↓↑

The inert gas neon (Z = l0), Ne, marks the end of the second period in the periodic table of elements.

In neon the L (n = 2) shell is complete; all its states are occupied by paired electrons.

Possessing no electrons outside a closed shell, the element neon is chemically inert.

The stability and low energy of a closed electronic configuration is evinced by the neon atom's relatively high first ionization potential of of 21.5eV.

Fluorine, F, the element before neon in the periodic table is short of just one electron to complete the L shell.

The fluorine atom's *electron afffinity* of -3.4 eV is further evidence of the low energy associated with a closed shell

Since there is no change in the shielding of the nuclear charge by the closed K shell as it increases from +3e in lithium to +10e in neon, the effective nuclear charge that acts on the outer L shell electrons increases.

Thus, the outer electrons are less tightly bound in lithium than in neon and the first ionization potential rises from 5.4eV in lithium to 21.5eV in neon.

The third and fourth periods commence with the elements sodium, Na, and potassium, K, respectively:

Na	**11 electrons**	$1s^2\ 2s^2\ 2p^6\ 3s^1$
K	**19 electrons**	$1s^2\ 2s^2\ 2p^6\ 3s^2\ 3p^6\ 4s^1$

Both sodium and potassium have low first ionization potentials; 5.2eV and 4.3eV respectively.

This is attributed to the screening of the nuclear charge by the closed inner shells, as in lithium.

Since the magnetic quantum number, m_l, can take five discrete values when $l = 2$, the d sub-level that becomes available in the M (n = 3) shell, can have a maximum population of ten electrons.

However, in multielectron atoms the energies of the electronic levels do not follow the same sequence as in hydrogen; they do not depend exclusively on the principal quantum number, n.

As Fig 3.30 shows, the energy of the 4s level in multielectron atoms is in fact lower than that of the 3d level.

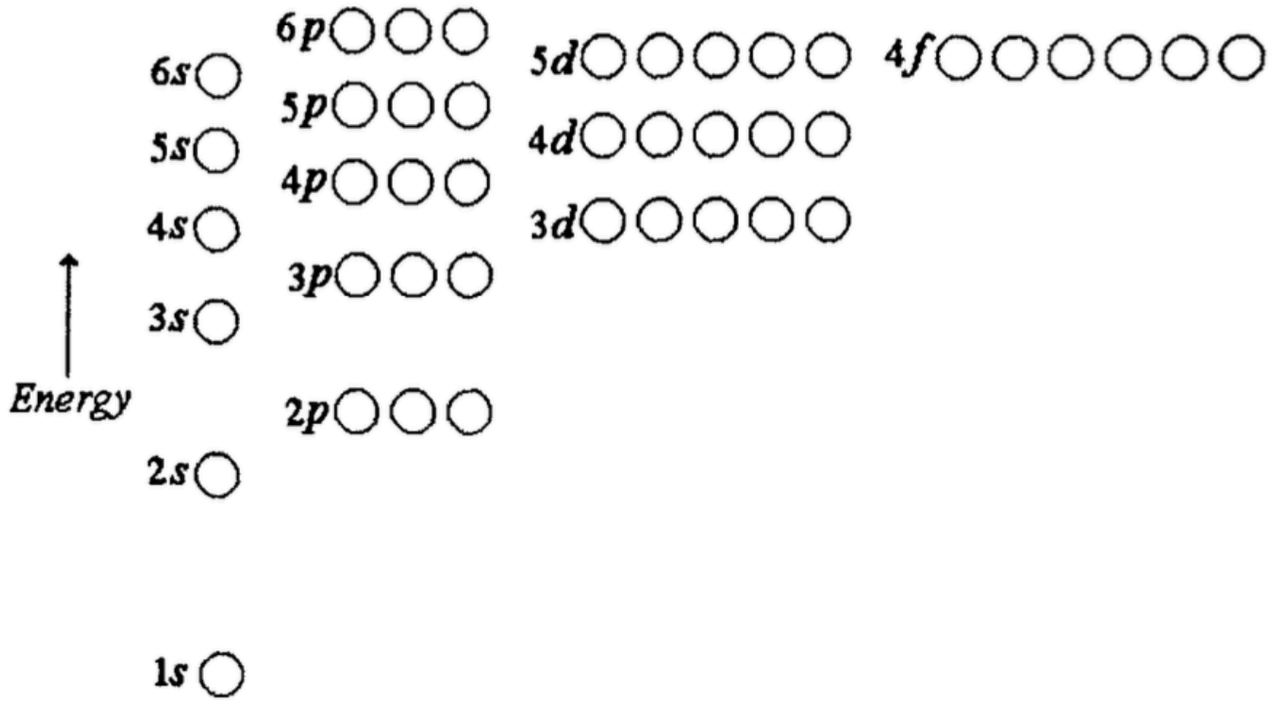


Fig 3.30 The sequence of energy levels in multi-electron atoms. In highly ionized species the order may differ due to the change in the effective nuclear charge. Each circle represents a quantum state that can be occupied by a pair of electrons possessing opposite spins.

Thus, electrons are introduced into the 4s level before they are assigned to the 3d level.

As a result, only the 3s and 3p levels are filled across the third period and it comprises just eight elements like the second period.

It is only in the fourth period that the 3d levels are progressively filled in the group of elements from scandium, Sc, to zinc, Zn.

These ten elements constitute the first row of the transition metals or d-block elements in the periodic table.

A similar situation arises in the N (n = 4) shell. Since the magnetic quantum number, m_l, can take seven discrete values when $l = 3$, the f sub-level that becomes available in this shell, can have a maximum population of fourteen electrons.

However, the energy of the 4f level is greater even than that of the 6s level and so it is only in the sixth period that the 4f levels are progressively filled in the group of elements from cerium, Ce, to lutetium, Lu, called the rare-earth or lanthanides.

As a result, the fifth period-the O (n = 5) shell-like the fourth period, contains eighteen elements of which ten are the second row of d-block elements.

The sixth period-the P (n = 6) shell-contains thirty-two elements.

These include the fourteen rare earths and the third row of d-block elements, apart from the two 6s and six 6p elements.

Since the chemical properties of an element are largely determined by its outermost electrons, the elements in each period of d-block elements are chemically similar; the difference between them is in the number of electrons they possess in the shell below that of their valence electrons.

The phenomenon is even more marked in the case of the rare-earth elements; they are almost chemically indistinguishable.

The only difference between them is in the number of electrons they possess in the 4f level, i.e., in a level that is two principal quantum numbers lower that that of their valence electrons.

The seventh period-the Q (n = 7) shell-is incomplete; there are just not enough elements to fill it.

The highest naturally occurring atomic number, Z, is 92, that of the element uranium, U.

A further seventeen elements-transuranic elements-have been manufactured by nuclear processes, all except plutonium, Pu, in just microscopic quantities.

The group of elements from thorium, Th, to lawrencium, Lw, constitute the actinides.

They are analogous to the lanthanides, differing from one another only in the number of 5f electrons they possess.

The electronic configurations of the 109 named naturally occurring and artificially manufactured elements are shown in Table 3.7.

Table 3.7: The electronic configuration of the elements.

			K	L		M			N				O				P		Q
			1*s*	2*s*	2*p*	3*s*	3*p*	3*d*	4*s*	4*p*	4*d*	4*f*	5*s*	5*p*	5*d*	5*f*	6*s*	6*p*	6*d*	7*s*
Period																				
1	1 Hydrogen	H	1																	
	2 Helium	He	2																	
2	3 Lithium	Li	2	1																
	4 Beryllium	Be	2	2																
	5 Boron	B	2	2	1															
	6 Carbon	C	2	2	2															
	7 Nitrogen	N	2	2	3															
	8 Oxygen	O	2	2	4															
	9 Fluorine	F	2	2	5															
	10 Neon	Ne	2	2	6															
3	11 Sodium	Na	2	2	6	1														
	12 Magnesium	Mg	2	2	6	2														
	13 Aluminium	Al	2	2	6	2	1													
	14 Silicon	Si	2	2	6	2	2													
	15 Phosphorus	P	2	2	6	2	3													
	16 Sulphur	S	2	2	6	2	4													
	17 Chlorine	Cl	2	2	6	2	5													
	18 Argon	Ar	2	2	6	2	6													
4	19 Potassium	K	2	2	6	2	6		1											
	20 Calcium	Ca	2	2	6	2	6		2											
	21 Scandium	Sc	2	2	6	2	6	1	2											
	22 Titanium	Ti	2	2	6	2	6	2	2											
	23 Vanadium	V	2	2	6	2	6	3	2											
	24 Chromium	Cr	2	2	6	2	6	5	1											
	25 Magnesium	Mn	2	2	6	2	6	5	2											
	26 Iron	Fe	2	2	6	2	6	6	2											
	27 Cobalt	Co	2	2	6	2	6	7	2											
	28 Nickel	Ni	2	2	6	2	6	8	2											
	29 Copper	Cu	2	2	6	2	6	10	1											
	30 Zinc	Zn	2	2	6	2	6	10	2											
	31 Gallium	Ga	2	2	6	2	6	10	2	1										
	32 Germanium	Ge	2	2	6	2	6	10	2	2										
	33 Arsenic	As	2	2	6	2	6	10	2	3										
	34 Selenium	Se	2	2	6	2	6	10	2	4										
	35 Bromine	Br	2	2	6	2	6	10	2	5										
	36 Krypton	Kr	2	2	6	2	6	10	2	6										
5	37 Rubidium	Rb	2	2	6	2	6	10	2	6			1							
	38 Strontium	Sr	2	2	6	2	6	10	2	6			2							
	39 Yttrium	Y	2	2	6	2	6	10	2	6	1		2							
	40 Zirconium	Zr	2	2	6	2	6	10	2	6	2		2							
	41 Niobium	Nb	2	2	6	2	6	10	2	6	4		1							
	42 Molybdenum	Mo	2	2	6	2	6	10	2	6	5		1							
	43 Technetium	Tc	2	2	6	2	6	10	2	6	5		2							
	44 Ruthenium	Ru	2	2	6	2	6	10	2	6	7		1							
	45 Rhodium	Rh	2	2	6	2	6	10	2	6	8		1							
	46 Palladium	Pd	2	2	6	2	6	10	2	6	10									
	47 Silver	Ag	2	2	6	2	6	10	2	6	10		1							
	48 Cadmium	Cd	2	2	6	2	6	10	2	6	10		2							
	49 Indium	In	2	2	6	2	6	10	2	6	10		2	1						
	50 Tin	Sn	2	2	6	2	6	10	2	6	10		2	2						
	51 Antimony	Sb	2	2	6	2	6	10	2	6	10		2	3						
	52 Tellurium	Te	2	2	6	2	6	10	2	6	10		2	4						
	53 Iodine	I	2	2	6	2	6	10	2	6	10		2	5						
	54 Xenon	Xe	2	2	6	2	6	10	2	6	10		2	6						

d-block elements

d-block elements

				K	L		M			N				O				P		Q
				1*s*	2*s*	2*p*	3*s*	3*p*	3*d*	4*s*	4*p*	4*d*	4*f*	5*s*	5*p*	5*d*	5*f*	6*s*	6*p*	6*d*	7*s*
Period																					
6	55	Caesium	Cs	2	2	6	2	6	10	2	6	10		2	6			1			
	56	Barium	Ba	2	2	6	2	6	10	2	6	10		2	6			2			
	57	Lanthanum	La	2	2	6	2	6	10	2	6	10		2	6	1		2			
	58	Cerium	Ce	2	2	6	2	6	10	2	6	10	2	2	6			2			
	59	Praseodymium	Pr	2	2	6	2	6	10	2	6	10	3	2	6			2			
	60	Neodymium	Nd	2	2	6	2	6	10	2	6	10	4	2	6			2			
	61	Promethium	Pm	2	2	6	2	6	10	2	6	10	5	2	6			2			
	62	Samarium	Sm	2	2	6	2	6	10	2	6	10	6	2	6			2			
	63	Europium	Eu	2	2	6	2	6	10	2	6	10	7	2	6			2			
	64	Gadolinium	Gd	2	2	6	2	6	10	2	6	10	7	2	6	1		2	Lanthanides		
	65	Terbium	Tb	2	2	6	2	6	10	2	6	10	9	2	6			2	(rare-		
	66	Dysprosium	Dy	2	2	6	2	6	10	2	6	10	10	2	6			2	earths)		
	67	Holmium	Ho	2	2	6	2	6	10	2	6	10	11	2	6			2			
	68	Erbium	Er	2	2	6	2	6	10	2	6	10	12	2	6			2			
	69	Thulium	Tm	2	2	6	2	6	10	2	6	10	13	2	6			2			
	70	Ytterbium	Yb	2	2	6	2	6	10	2	6	10	14	2	6			2			
	71	Lutetium	Lu	2	2	6	2	6	10	2	6	10	14	2	6	1		2			
	72	Hafnium	Hf	2	2	6	2	6	10	2	6	10	13	2	6	2		2			
	73	Tantalum	Ta	2	2	6	2	6	10	2	6	10	14	2	6	3		2			
	74	Tungsten	W	2	2	6	2	6	10	2	6	10	14	2	6	4		2			
	75	Rhenium	Re	2	2	6	2	6	10	2	6	10	13	2	6	5		2	d-block		
	76	Osmium	Os	2	2	6	2	6	10	2	6	10	14	2	6	6		2	elements		
	77	Iridium	Ir	2	2	6	2	6	10	2	6	10	14	2	6	7		2			
	78	Platinum	Pt	2	2	6	2	6	10	2	6	10	13	2	6	9		1			
	79	Gold	Au	2	2	6	2	6	10	2	6	10	14	2	6	10		1			
	80	Mercury	Hg	2	2	6	2	6	10	2	6	10	14	2	6	10		2			
	81	Thallium	Tl	2	2	6	2	6	10	2	6	10	14	2	6	10		2	1		
	82	Lead	Pb	2	2	6	2	6	10	2	6	10	14	2	6	10		2	2		
	83	Bismuth	Bi	2	2	6	2	6	10	2	6	10	14	2	6	10		2	3		
	84	Polonium	Po	2	2	6	2	6	10	2	6	10	14	2	6	10		2	4		
	85	Astatine	At	2	2	6	2	6	10	2	6	10	14	2	6	10		2	5		
	86	Radon	Rn	2	2	6	2	6	10	2	6	10	14	2	6	10		2	6		
7	87	Francium	Fr	2	2	6	2	6	10	2	6	10	14	2	6	10		2	6		1
	88	Radium	Ra	2	2	6	2	6	10	2	6	10	14	2	6	10		2	6		2
	89	Actinium	Ac	2	2	6	2	6	10	2	6	10	14	2	6	10		2	6	1	2
	90	Thorium	Th	2	2	6	2	6	10	2	6	10	14	2	6	10		2	6	2	2
	91	Protactinium	Pa	2	2	6	2	6	10	2	6	10	14	2	6	10	2	2	6	1	2
	92	Uranium	U	2	2	6	2	6	10	2	6	10	14	2	6	10	3	2	6	1	2
	93	Neptunium	Np	2	2	6	2	6	10	2	6	10	14	2	6	10	4	2	6	1	2
	94	Plutonium	Pu	2	2	6	2	6	10	2	6	10	14	2	6	10	5	2	6	1	2
	95	Americium	Am	2	2	6	2	6	10	2	6	10	14	2	6	10	6	2	6	1	2
	96	Curium	Cm	2	2	6	2	6	10	2	6	10	14	2	6	10	7	2	6	1	2
	97	Berkelium	Bk	2	2	6	2	6	10	2	6	10	14	2	6	10	8	2	6	1	2
	98	Californium	Cf	2	2	6	2	6	10	2	6	10	14	2	6	10	10	2	6		2
	99	Einsteinium	Es	2	2	6	2	6	10	2	6	10	14	2	6	10	11	2	6		2
	100	Fermium	Fm	2	2	6	2	6	10	2	6	10	14	2	6	10	12	2	6		2
	101	Mendelevium	Md	2	2	6	2	6	10	2	6	10	14	2	6	10	13	2	6		2
	102	Nobelium	No	2	2	6	2	6	10	2	6	10	14	2	6	10	14	2	6		2
	103	Lawrencium	Lw	2	2	6	2	6	10	2	6	10	14	2	6	10	14	2	6	1	2
	104	Rutherfordium	Rf	2	2	6	2	6	10	2	6	10	14	2	6	10	14	2	6	2	2
	105	Dubnium	Db	2	2	6	2	6	10	2	6	10	14	2	6	10	14	2	6	3	2
	106	Seaborgium	Sg	2	2	6	2	6	10	2	6	10	14	2	6	10	14	2	6	4	2
	107	Bohrium	Bh	2	2	6	2	6	10	2	6	10	14	2	6	10	14	2	6	5	2
	108	Hassium	Hs	2	2	6	2	6	10	2	6	10	14	2	6	10	14	2	6	6	2
	109	Meitnerium	Mt	2	2	6	2	6	10	2	6	10	14	2	6	10	14	2	6	7	2

3.4.5 Spin, Identical Particles and Pauli's Principle

In classical mechanics, intrinsically identical particles, such as all the red balls on a snooker table, can be distinguished from one another by means of their individual trajectories.

Should the particles collide, their trajectories after the collision can be predicted, in principle if not always in practice, from the laws of classical mechanics.

However, in quantum mechanics, particles cannot be assigned trajectories; they cannot be tracked. Quantum mechanics pictures collisions in terms of the overlapping (superposition) of the particles' wave functions. On this view, we can never be sure, after the collision, which particle is which.

There is no way of knowing exactly how the composite wave function formed at the collision 'unwinds' itself afterwards.

How then, if at all, can intrinsically identical particles be distinguished from one another in quantum mechanics?

In order to investigate this question, let us suppose that following a collision between two identical minute particles, P_1 and P_2, one appears at A and the other at B (Fig. 3.31).

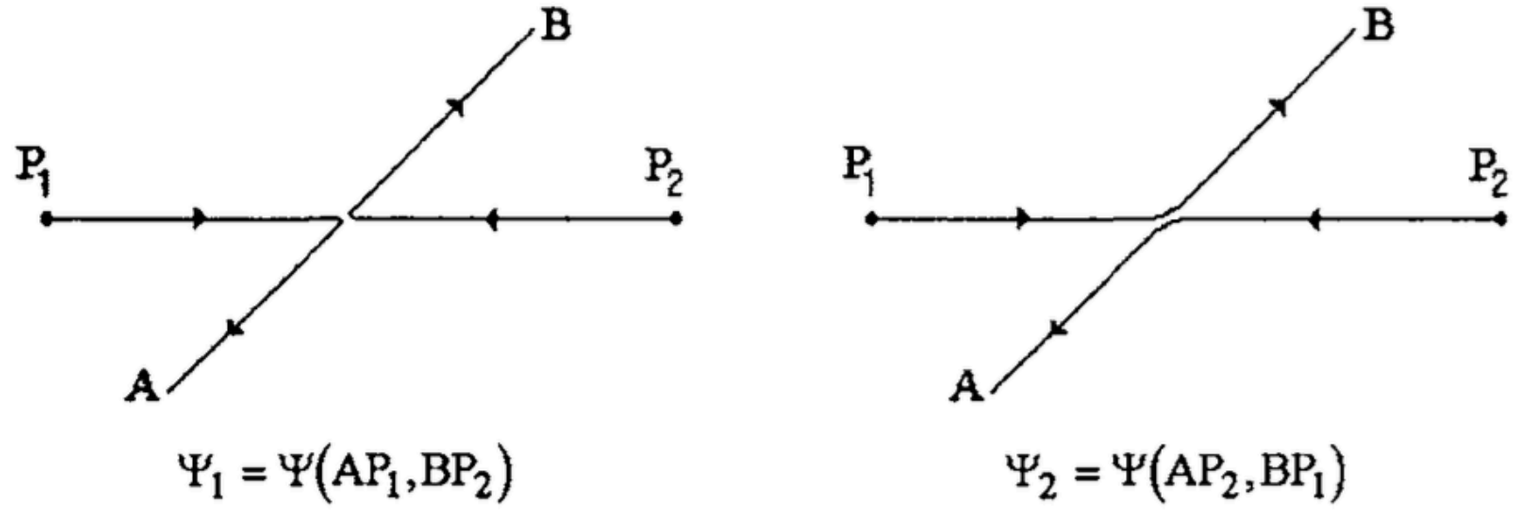


Fig.3.31 A collision between two intrinsically identical particles after which one is found at point A and another at point B. One possibility is that particle 1 arrives at point A and particle 2 at point B, the other is that particle 1 arrives at point B and particle 2 at point A. Each of these two possibilities is characterized by its own wave function.

There are two different ways in which this can occur:

1. **The particle that arrives at point A after the collision is particle P_1 and that which arrives at point B is P_2;**
2. **The particle that arrives at point A after the collision is particle P_2 and that which arrives at point B is P_1;**

In accordance with the rules for describing events in terms of wave functions, the wave function, Ψ, that describes the collision is:

$$\Psi = \Psi_1 + \Psi_2 \tag{3.59}$$

where $\Psi_1 = \Psi(AP_1, BP_2)$ and $\Psi_2 = \Psi(AP_2, BP_1)$ are the wave functions of the two alternative ways the event can occur.

Both wave functions express the possibility of finding one particle at A and another at B after the collision

Because the particles are identical (exchangeable), the two possible outcomes of the collision, (AP_1, BP_2) and (AP_2, BP_1) , are indistinguishable and equally probable.

Consequently, their probability densities must be equal:

$$\left|\Psi(AP_1, BP_2)\right|^2 = \left|\Psi(AP_2, BP_1)\right|^2 \tag{3.60}$$

There are only two ways that the wave functions, $\Psi_1 = \Psi(AP_1, BP_2)$ and $\Psi_2 = \Psi(AP_2, BP_1)$ can be related so as to satisfy equation (3.82)

(i) The two wave functions are in phase with each other:

$$\Psi(AP_1, BP_2) = \Psi(AP_2, BP_1) \tag{3.61}$$

In this case the wave function, Ψ_S, that describes the collision is

$$\Psi_S = \frac{1}{\sqrt{2}}\left[\Psi(AP_1, BP_2) + \Psi(AP_2, BP_1)\right] \tag{3.62}$$

This type of function is called a *symmetric wave function*.

(ii) The two wave functions are out of phase with each other, there being a phase difference of π between them:

$$\Psi(AP_1, BP_2) = -\Psi(AP_2, BP_1) \tag{3.63}$$

In this case the wave function, Ψ_A, describing the collision is

$$\Psi_A = \frac{1}{\sqrt{2}}\left[\Psi(AP_1, BP_2) - \Psi(AP_2, BP_1)\right] \tag{3.64}$$

This type of function is called an *anti-symmetric wave function*.

In both cases the factor $1/\sqrt{2}$ is required to normalize the wave function.

Whether the wave function describing an event is symmetric or anti-symmetric depends on the nature of the particles.

Those particles for which the appropriate wave functions are symmetric, like equation (3.62), are called *bosons*; those for which they are anti-symmetric, like equation (3.64), are called *fermions*.

Is there any chance that both particles might actually finish up at the same point, say at the point A, after the collision?

Strange as this may appear to those accustomed to picturing all collisions between particles in terms of snooker or billiards, in quantum mechanics there is no a priori reason why this should not happen; the overlapping and subsequent unwinding of the wave functions does not of necessity preclude this outcome.

For example, the symmetric wave function, Ψ_S, does not collapse, i.e., it does not reduce to zero ($\Psi_S \neq 0$) , if 'A' is substituted in place of 'B' in equation (3.84):

$$\Psi_S = \frac{1}{\sqrt{2}}\left[\Psi(AP_1, AP_2) + \Psi(AP_2, AP_1)\right] \tag{3.65}$$

And so, strange as it may appear, there is a certain probability that after a collision between two bosons they will both turn up at the same place.

Putting this another way, there is nothing to prevent two or even more bosons from simultaneously occupying the same state.

On the other hand, the anti-symmetric, Ψ_A, wave function does collapse if 'A' is substituted in place of 'B' in equation (3.82):

$$\Psi_A = \frac{1}{\sqrt{2}}\left[\Psi(AP_1, AP_2) - \Psi(AP_2, AP_1)\right] = 0 \tag{3.66}$$

Thus, there is a zero probability of the event taking place in this way; two identical fermions can never simultaneously occupy the same state.

However, this is exactly the import of Pauli's principle:

In a system containing a number of electrons, no two electrons can occupy the same state; each electron must have a different set of quantum numbers.

Let us suppose that two electrons collide.

According to Pauli's principle, they cannot both finish up in the same state.

This is equivalent to saying that the wave function which describes the collision must be an anti-symmetric wave function, Ψ_A.

It follows, that electrons are fermions and that Pauli's principle can be expressed alternatively as requiring that the wave function, that describes the electrons in a system, must be anti-symmetric.

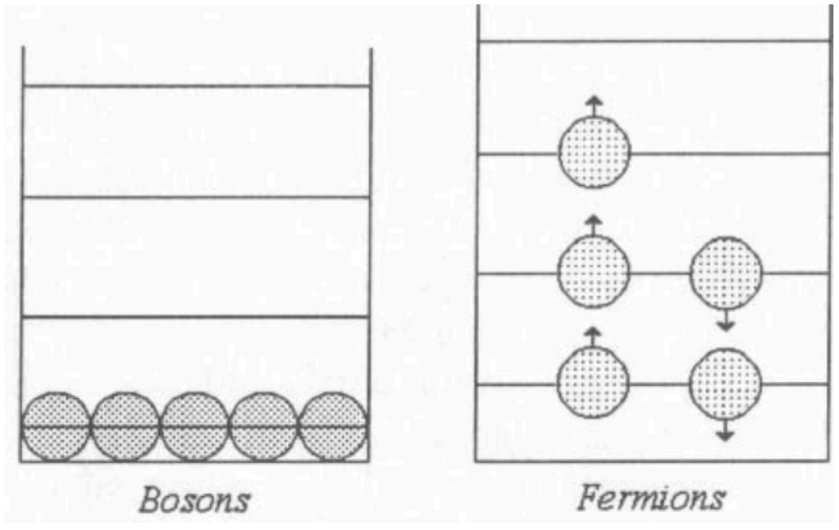
In general, particles whose spin is an integer (0,1, 2, ...) are bosons: photons, α particles and helium atoms are bosons.

Pauli's exclusion principle does not apply to them, and so there is no restriction on the number of bosons that can occupy the ground state.

On the other hand, particles whose spin is an odd multiple of a half (1/2, 3/2, 5/2,…) are fermions: electrons, protons and neutrons are fermions. Pauli's exclusion principle does apply to them, and the number of fermions that can occupy any state is restricted.

Thus, in multi-electron atoms, higher electronic energy states must also be occupied (Fig. 3.32)

Fig. 3.32 Bosons and fermions in a quantum box. Each system tends to minimum energy. The system of bosons can achieve this by placing all the bosons in the ground state. However, the system of fermions can allocate a pair of particles at most to each of the energy levels.



In order to determine the distribution of a system's particles between its energy levels, different statistical functions must be used depending on the type of particle: classical particles, bosons or fermions.

Since, under ordinary conditions, the average distance between the atoms and molecules of a gas is greater than the wavelength of the matter wave associated with them, they can be distinguished by means of their trajectories.

It follows that they can be described by classical physics.

It can be shown, that the probability, $f(E_i)$, of a classical particle occupying the E_i, energy level at an absolute temperature T is given by:

$$f(E_i) = \left(Ae^{E_i/kT} \right)^{-1} \tag{3.67}$$

where the parameter A depends on the total number of particles in the system.

In such systems there are no restrictions on the number of particles that can occupy a particular state, nor does the presence of a particle in a particular state have any effect on the other particles in the system.

The statistical function which describes a system of bosons is called *Bose-Einstein statistics*.

The probability, $f(E_i)_{Boson}$ of a boson, occupying the E_i energy level at an absolute temperature T is given by:

$$f(E_i)_{Boson} = \left(Be^{E_i/kT} - 1 \right)^{-1} \tag{3.68}$$

where the parameter B depends on the nature of the system.

In a system of bosons there are no restrictions on the number of particles that can occupy a particular state.

The presence of a particle in a particular state even increases the probability that other particles will be found in the same state.

Consequently, there is a tendency for the lowest energy states to be occupied, and at very low temperatures most of the particles will be found in the ground state.

The photon gas that fills a constant-temperature cavity is an example of a system of bosons; the spectrum of black-body radiation can be explained in terms of the distribution of the photons in the cavity between the various energy levels as required by the Bose-Einstein function.

Helium atoms can also behave as bosons; liquid helium, at temperatures below 2.18K, exhibits strange properties that stem from this.

At this very low temperature, $f(E_i)_{Boson}$ —> 0 for all but the lowest energy level.

The higher levels are in effect completely empty and all the helium atoms 'condense' into the lowest energy level where they cooperate with one another.

This gives the liquid strange properties.

It shows no viscosity and flows out of an open container by spontaneously climbing up its inside walls.

The statistical function which describes a system of fermions is called *Fermi-Dirac statistics*.

In a system of fermions only one particle, and no more, can occupy each state, and so the presence of a particle in a particular state prevents other particles from being in that state.

The probability, $f(E_i)_{Fermion}$, of a fermion occupying the E_i energy level at an absolute temperature T is given by:

$$f(E_i)_{Fermion} = \left(Ce^{E_i/kT} + 1 \right)^{-1} \tag{3.69}$$

where the parameter C depends on the nature and temperature of the system.

The electron gas in metal conductors is an example of a system of fermions.

3.4.6 Total Spin

Whereas in the elucidation of the structure of the periodic table we considered each atomic electron individually, it is often useful, in other instances, to regard the constituent electrons of an atom or molecule as a group or a whole.

The key to this approach is the concept of *total spin*.

The total intrinsic angular momentum, i.e., the total spin, **S**, of an atom (or molecule) is given by the vector sum of the spins of all its electrons:

$$\mathbf{S} = \mathbf{S}_1 + \mathbf{S}_2 + \mathbf{S}_3 + \ldots \tag{3.70}$$

where the eigenvalues of the square, S^2, of its magnitude are given by:

$$S^2 = S(S+1)\hbar^2 \tag{3.71}$$

S is called the *total spin quantum number*.

The component, S_z, of the total spin parallel to an arbitrarily chosen Z-axis is given by

$$S_Z = M_S\hbar \tag{3.72}$$

where the quantum number M_S is equal to the sum of the spin magnetic quantum numbers, m_{s1}, m_{S2},… m_{sN}, …… , of the N constituent electrons of the atom or molecule:

$$M_S = m_{s_1} + m_{s_2} + \ldots + m_{s_N} \tag{3.73}$$

The total spin quantum number, S, is subject to the following three rules:

1. a) **If the number, N, of the constituent electrons is even, the total spin quantum number S is an integer: $S = 0, 1, 2, \ldots$;**

   b) **if N is an odd number, the total spin quantum number S is an odd multiple of a half: $S = \frac{1}{2}, \frac{3}{2}, \frac{5}{2} \ldots$.**

2. **Because of Pauli's principle, the total spin quantum number of a closed shell or sub-shell, i.e., one that contains its full quota of electrons, must be zero: $S_{Closed\ shell} = 0$.**

3. **For any number, N', of electrons outside a closed shell, the highest possible value for S is $N'/2$.**

It follows, that the only possible total spin quantum numbers for a pair of electrons outside a closed shell (N' = 2) are S = 1 or S = 0.

In those cases where S = 0, the spins of the two electrons are anti-parallel, i.e., one is directed upwards and the other downwards; their resultant is then zero.

This is called a singlet state, since it allows for only one possible value for the quantum number M_S, namely, $M_S = 0$ (Fig 3.33a).

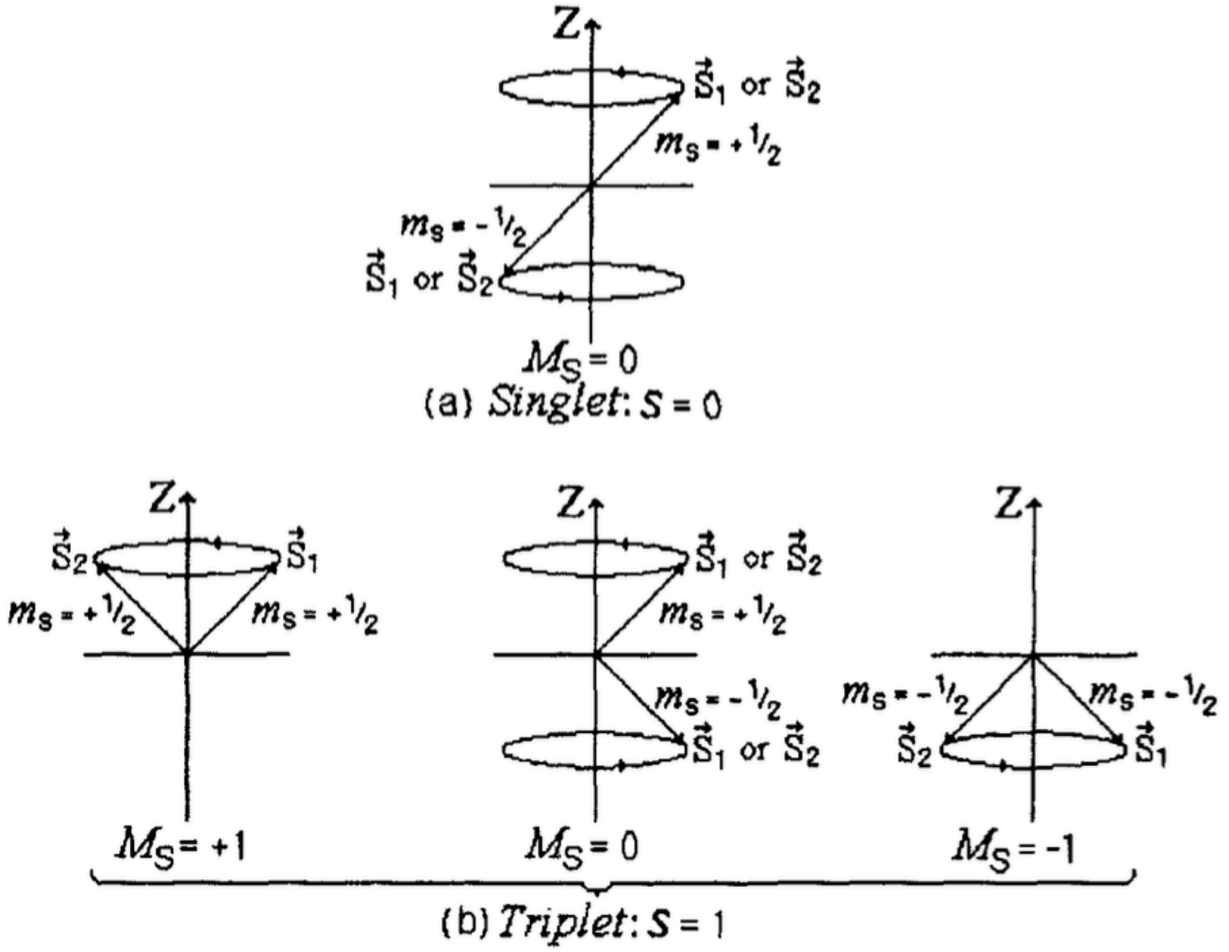


Fig.3.33 The spin states of a pair of electrons. (a) S = 0; singlet state. The spins of the two electrons are anti-parallel. (b) S = 1; triplet state. In the $M_S = 0$ state the spins of the electrons will be anti-parallel, in the M_S = -1 and $M_S = 1$ states the electron spins will be parallel.

When $S = 1$, the possible values of M_S are -1, 0, and +l; this is called a triplet state because it allows for three different values of M_S.

In the $S = 1$, $M_S = 0$ state the electron spins are anti-parallel; in the S=1, M_S =-1 and S=1, M_S =+1 states they are parallel.

Thus, in an external magnetic field, each $S = 1$ energy level splits into three states which show up as fine structure in the spectrum (Fig. 3.33b).

The total spin function of the pair of electrons comprising the singlet and triplet states can be expressed in terms of the one electron spin wave functions, $\chi_{1/2}$ or $\chi_{-1/2}$, that correspond, respectively, to the $m_s = +1/2$ and $m_s = -1/2$ states.

Since the electrons are indistinguishable, these wave functions can be equally ascribed to both, such that those of the one can be designated $\chi_{1/2}(e_1)$or $\chi_{-1/2}(e_2)$ and those of the other $\chi_{1/2}(e_2)$or $\chi_{-1/2}(e_1)$.

Whether the total spin function of the pair of electrons is symmetric, χ_S, or anti-symmetric, χ_A, depends on how their individual spin functions are combined.

In the singlet state the spins of the pair of electrons are always opposed. Hence, the total spin function of the singlet state must be anti-symmetric:

$$\chi_A = \frac{1}{\sqrt{2}}\left[\chi_{+\frac{1}{2}}(e_1)\chi_{-\frac{1}{2}}(e_2) - \chi_{+\frac{1}{2}}(e_2)\chi_{-\frac{1}{2}}(e_1)\right] \quad M_S = 0 \tag{3.74}$$

In the triplet state the three possible spin wave functions are all symmetric:

$$\chi_S = \begin{cases} \chi_{+\frac{1}{2}}(e_1)\chi_{+\frac{1}{2}}(e_2) & M_S = +1 \\ \frac{1}{\sqrt{2}}\left[\chi_{+\frac{1}{2}}(e_1)\chi_{-\frac{1}{2}}(e_2) + \chi_{+\frac{1}{2}}(e_2)\chi_{-\frac{1}{2}}(e_1)\right] & M_S = 0 \\ \chi_{-\frac{1}{2}}(e_1)\chi_{-\frac{1}{2}}(e_2) & M_S = -1 \end{cases} \tag{3.75}$$

3.4.7 The Energy Levels in Multi-electron Atoms

Whereas the total energy of the single electron in a hydrogen atom depends just on its kinetic energy and the potential energy of its attractive interaction with the nucleus, that of the electrons in multi-electron atoms is also affected by the repulsive electrostatic interaction between the electrons themselves.

This repulsive interaction between the electrons raises the total energy of the atom.

The magnitude of this repulsive interaction depends on two factors:

(i) the atom's total spin quantum number, S;

(ii) their spatial separation (the average distance between them).

Each of these factors can be represented by a wave function: the former by the total spin wave function, χ, the latter by the wave function, ϕ, usually referred to as the spatial or orbital wave function.

These wave functions can be symmetric, χ_S and ϕ_A or anti-symmetric, χ_A and ϕ_S,.

The total spin function will be symmetric or anti-symmetric depending on whether the total spin state is a triplet or singlet.

As regards the spatial wave function (the orbital), it will be symmetric if the electrons can share the same position in space and anti-symmetric if they cannot.

In accordance with the rules for describing quantum-mechanical systems, the overall wave function that describes the atomic electrons is given by the product, $\Psi = \phi$ x χ.

In principle, this gives four possibilities for the overall wave function:

	Symmetric orbital, ϕ_S	Anti-symmetric orbital, ϕ_A
Singlet ($S = 0$)	$\psi_A = \phi_S \times \chi_A$	$\psi_S = \phi_A \times \chi_A$
Triplet ($S = 1$)	$\psi_S = \phi_S \times \chi_S$	$\psi_A = \phi_A \times \chi_S$

The overall wave function wiII be symmetric, Ψ_S, if both factors are symmetric or both are anti-symmetric; it will be anti-symmetric, Ψ_A, if one is symmetric and the other anti-symmetric.

Because they are fermions, the overall wave function that describes the repulsive interaction between the electrons in a multielectron atom must be an anti-symmetric wavefunction.

Clearly, in practice the only two possibilities are:

$$\psi_{Triplet} = \phi_A \times \chi_S \tag{3.76}$$

and

$$\psi_{Singlet} = \phi_S \times \chi_A \tag{3.77}$$

The simplest multi-electron atom is helium.

In its ground state, the two electrons possessed by the atom share three identical quantum numbers, n = 1, $l = 0$ and $m_l = 0$.

The electrons constitute a closed shell and in these circumstances (rule 2 above), Pauli's principle does not permit a spin of S = 1.

Their fourth quantum number-the spin magnetic quantum number-must distinguish between them: that of the one must be $m_s = +1/2$ and that of the other $m_s = -1/2$.

As a result, the total spin magnetic quantum number must at all times be zero:

$$M_S = m_{s1} + m_{s2} = ½ + (-½) = 0 \qquad (3.78)$$

This restricts the ground state to the singlet state in which S = 0.

No such restriction applies to the excited states in the helium atom.

Two distinct groups of excited states exist -singlet and triplet.

In a sense, we may regard each group as comprising a different gas:

* *Parahelium* for which $S = 0$;
* *Orthohelium* for which $S = 1$.

In general, the energy levels of orthohelium (the triplet states of helium) are found to be lower than the corresponding parahelium (singlet state) energy levels (Fig 3.34):

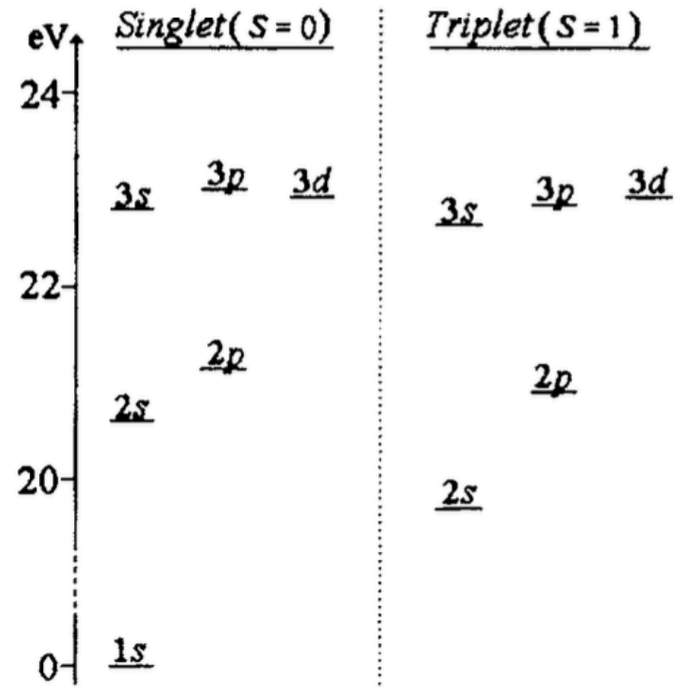


Fig 3.34 The energy level diagram for helium for the principal quantum numbers n = 1, 2 and 3 . The energies are given relative to the ground state. One of the atom's electrons is in the pound state and the other in the indicated ns, np or nd excited state. The energy levels of the singlet states (parahelium) are generally higher than those of the corresponding triplet (orthohelium) states.

This can be explained as follows:

The probability of the two helium electrons being found together at the same point in space is lower in the triplet state where the spatial function is anti-symmetric than in the singlet state where it is symmetric.

Hence, on average, the electrons will be further apart in the triplet state and the repulsive electrostatic interaction between them will be correspondingly weaker.

Consequently, the repulsive electrostatic interaction will raise the energy of the system less in the triplet state than in the singlet state.

Whereas transitions that conform to the selection rules $\Delta l = \pm 1$ and $\Delta m_l = 0, \pm 1$ are permitted and do occur within each helium state-singlet or triplet, spectral transitions between the triplet state and the singlet state do not occur under usual circumstances.

Such transitions would require a spin rearrangement and so their probability is very low.

On this basis we can formulate a further selection rule for spectral transitions:

Transitions obey the selection rule $\Delta S = 0$.

This rule always applies in systems where the potential energy does not explicitly include the spin.

Because a fall from the lowest triplet states in helium violates this selection rule, these states are metastable.

Unless a collision occurs, a helium atom can remain in these states for a relatively long time -a second or more- before radiating.

In multi-electron atoms that possess two valence electrons, such as beryllium, magnesium and calcium, the situation is very similar to that in helium.

Their energy levels also fall into two groups, singlet (S = 0) and triplet (S = 1); these correspond to the excitation of just one of the valence electrons.

Transitions within the two groups that obey the selection rules $\Delta l = \pm 1$ and $\Delta m_l = 0, \pm 1$ readily occur; those between the groups that would violate the rule $\Delta S = 0$ are very weak.

Electrostatic interactions are not the only ones that act upon the components of an atom; magnetic interactions also affect them.

An example of this is the interaction of the spin magnetic moment of the electron with the magnetic field induced by the nucleus-spin-orbit coupling.

The spectral line splitting that results from this magnetic interaction is only a minor effect in the atoms of elements at the beginning of the periodic table, especially of those which constitute organic material -hydrogen, carbon, nitrogen and oxygen.

However, as the atomic number increases, so does the importance of this magnetic interaction.

The total spin, S, of the electrons outside the closed shells then ceases to be a good quantum number with which to characterize the energy levels, and the selection rule $\Delta S = 0$ is no longer found to hold.

3.4.8 Total Spin and the Energy Levels in Molecules

Early attempts to explain the nature of the chemical bond, such as the dualistic theory, had supposed all bonds to be polar and to involve the transfer of charge from one of the linked atoms to the other.

According to this theory, each bond between two atoms represents the transfer of a valency electron from one atom to the other.

As a result, one of the atoms becomes positively charged, the other negatively charged.

This creates an electrical dipole whose ends are attracted to one another.

Although this approach succeeded in explaining the structure of ionic crystals, e.g. $Na^{+}CI^{-}$, it proved incompetent to account for the behavior of relatively non-polar compounds such as organic compounds.

Stark first put forward the improved idea that the valency electrons released by the linked atoms attract their positively charged components concurrently.

He regarded these electrons as being situated somewhere in the space between the atoms (Thomson-type atoms) and bonding them electrostatically.

In an extension of this idea and of his own atomic model, Bohr imagined the chemical bond to be a ring of circulating electrons in an orbit whose plane was perpendicular to the line joining the atomic centers.

In 1907, Thomson had conjectured that atoms having the structure he had postulated might be bonded by electric forces without polarization occurring.

He suggested, that in homonuclear molecules such H_2 or O_2 the atoms overlapped with their electrons situated symmetrically in the region of overlap.

In this description, the spheres of positive charge were held together by the localization and concentration of their electrons between them.

This model did not require a molecular dipole to bond the atoms.

On the contrary, it postulated the revolutionary notion, that bonding arises from a localization of the charges within each atom.

The modem conception of the chemical bond was first proposed by Lewis in 1916.

> ***The simple assumption [is] that the chemical bond is at all times and in all molecules merely a pair of electrons held jointly by two atoms.***

However, an explanation of how two negatively charged entities could co-operate to create a bond despite their mutual electrostatic repulsion, had to await the development of quantum mechanics.

The key, in this case too, is the concept of total spin.

In contrast to the situation in atoms, where electrons in the triplet state generally have the lower energy levels, the study of molecules shows that the electrons that constitute covalent bonds have a lower energy in the singlet state than in the triplet state.

It follows, that the wave function of a covalent bond has to be of the form

$$\Psi_{CovalentBond} = \chi_A \times \phi_S \qquad 3.79$$

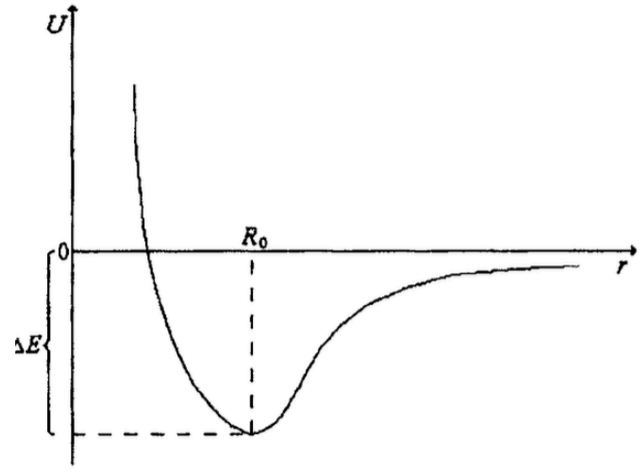
i.e., the spin wave function of the electrons constituting the covalent bond is anti-symmetric, χ_A, and the spatial wave function (orbital) is symmetric, ϕ_S.

In this state, the spatial wave functions (the orbitals) of the electrons overlap, a configuration that increases the electrostatic energy of repulsion between them.

However, covalent bond formation is always accompanied by the establishment of a minimum energy equilibrium (Fig 3.35).

How does the system compensate for the increased energy of repulsion between the electrons and achieve the overall reduction in its energy?

Fig 3.35 The potential energy diagram of bond formation. As the atoms approach, the potential energy, U, falls. It reaches its minimum value at the internuclear distance R_0 after which it rises once more as the repulsion between the nuclei becomes dominant. The energy released in the process, ΔE, is the bond energy.

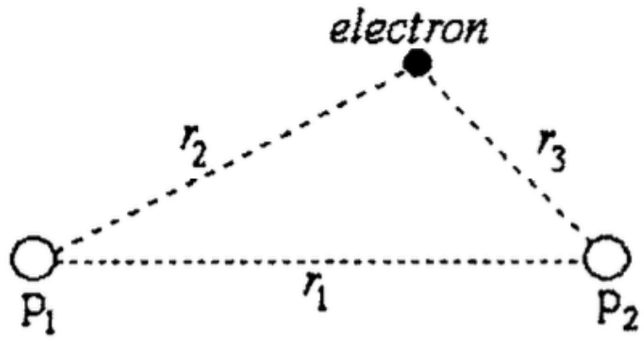


The effects that serve to reduce the system's energy when a bond is formed arise from the fact that each electron comprising the bond is attracted by the two nuclei of the bonded atoms and not just by the nucleus of its 'parent' atom.

These effects can be interpreted as having a Coulombic component and a purely quantum component.

The simplest system in which an electron is shared by two nuclei is the molecule H_2^+ formed when a hydrogen atom captures a proton (Fig 3.36).

Fig 3.36 The H_2^+ molecule formed when a hydrogen atom captures a proton. The two protons, p_1 and p_2, each carry a charge +e. The electron carries a charge -e.



The formal procedure for establishing this molecule's stationary states is:

(i) to find the expression for its potential energy, *U*;

(ii) to substitute this expression in the appropriate Schrödinger equation;

(iii) to solve the resulting equation.

When the hydrogen atom and the proton are far apart, the system's energy will be just that of the hydrogen atom, i.e., -13.6eV.

However, as they approach, the particles comprising the hydrogen atom and the second proton come under the influence of each others electric fields and the following additional electrostatic interactions arise, each with its own potential energy:

* **Between the nucleus of the hydrogen atom (its proton) and the second proton with a positive potential energy,** $U_{p-p} = ke^2/r_1$;
* **Between the electron and each proton with an overall negative potential energy,** $U_{p-e} = -\left(ke^2/r_2 + ke^2/r_3\right)$.

The total potential energy of these interactions is

$$\begin{aligned} U &= U_{p-p} + U_{p-e} \\ &= ke^2\left(1/r_1 - 1/r_2 - 1/r_3\right) \end{aligned} \qquad 3.80$$

This is the expression to be substituted in the Schrodinger equation.

The solution of the resulting equation is complex and beyond the scope of this class.

We can however, once again, use our physical intuition in order to obtain an understanding of the solution and of the result it gives for the energy of the molecule's ground state.

Though it originally belonged to the hydrogen atom, the electron loses this identity in the formation of the H_2^+ molecule.

Like the elusive particles that appear to pass simultaneously through both slits in experiments like that carried out by Young, no definite trajectory can be assigned to the electron.

It may even appear to be at two different places at the same time, orbiting both nuclei simultaneously.

And likewise, its spatial wave function can be viewed as arising from the superposition of a number of base states, $\phi = \phi_1 + \phi_2 + \phi_3 + ...$, with the corresponding probability density, $|\phi|^2$, involving interference terms in addition to the probability amplitudes of the base states.

It transpires, that most of the energy released when a covalent bond forms is associated with the interference terms that appear in the probability density.

This energy, whose origin is a pure quantum effect, is termed *exchange energy*.

Let two base states of the electron in the H_2^+ molecule be described by the spatial wave functions ϕ_1 and ϕ_2 (Fig 3.37).

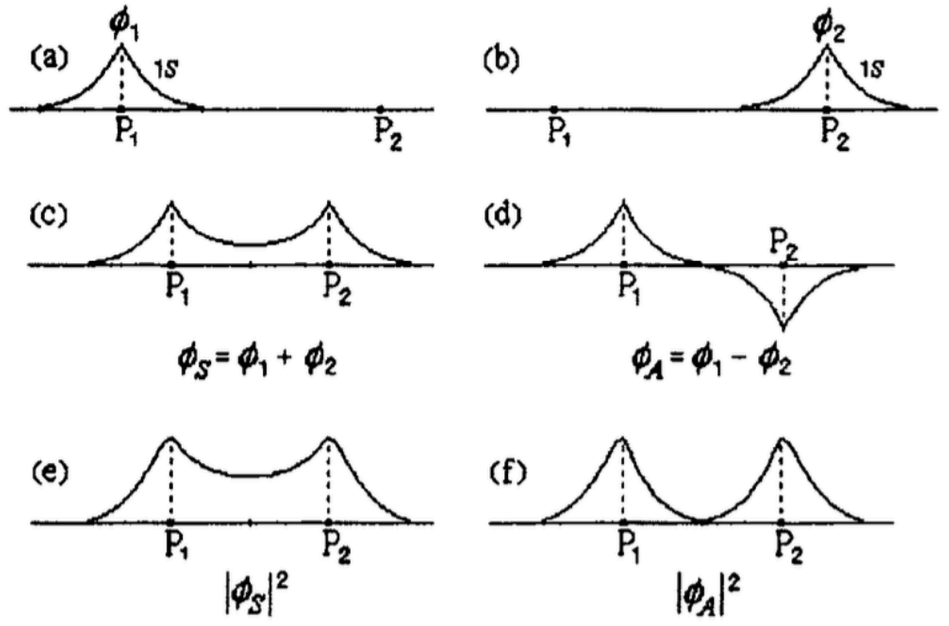


Fig. 3.37 A system of two protons and one electron. Two possible base states of the system are represented by (a) and (b), which show, respectively, the combination of the hydrogen atom P_1 with the proton P_2 and the combination of the proton P_1 with the hydrogen atom P_2. The spatial distribution of the electron in these states is described, respectively, by the wave functions, ϕ_1 and ϕ_2. In figures (c) to the proximity of the protons corresponds to the situation in the H_2^+ molecule. Figure (c) shows the symmetric spatial wave function (molecular orbital) $\phi_S = \phi_1 + \phi_2$; figure(d) the anti-symmetric wave function $\phi_A = \phi_1 - \phi_2$. Figures (e) and (f) show the corresponding probability densities, $|\phi_S|^2$ and $|\phi_A|^2$.

The effect of the interference term on the probability density of the state formed by the superposition of ϕ_1 and ϕ_2 depends on whether the wave functions of these base states are in phase with one another or out of phase.

If the wave functions are in phase, the resulting *molecular orbital*, ϕ_S, is symmetric.

If they are out of phase, the resulting orbital ϕ_A, will be anti-symmetric.

Ignoring the normalization requirements, this gives:

	Wavefunction	Probability Density
In phase:	$\phi_S = \phi_1 + \phi_2$	$\|\phi_S\|^2 = \phi_1^2 + \phi_2^2 + 2\phi_1\phi_2$
Out of phase:	$\phi_A = \phi_1 - \phi_2$	$\|\phi_A\|^2 = \phi_1^2 + \phi_2^2 - 2\phi_1\phi_2$

In general, the average (expectation) value of the electron's potential energy depends on the probability density of the spatial wave function, $|\phi|^2$:

$$\begin{aligned}\left\langle U_{p-e}\right\rangle &\propto \int U_{p-e}|\phi|^2\,dr \\ &= \int U_{p-e}\phi_1^2\,dr + \int U_{p-e}\phi_2^2\,dr \pm \int U_{p-e}2\phi_1\phi_2\,dr\end{aligned} \tag{3.81}$$

The contribution of the Coulombic interaction between the electron and the protons to the potential energy depends on the first two terms in this expression, those in which the probability densities of the two base states, ϕ^2_1 and ϕ^2_2 , appear.

$$\left(U_{p-e}\right)_{Coulomb} \propto \int U_{p-e}\phi_1^2\,dr + \int U_{p-e}\phi_2^2\,dr \tag{3.82}$$

The contribution of the exchange energy depends on the third term in equation (3.81), that in which the interference factor, 24h, appears. (3.83)

$$\left(U_{p-e}\right)_{Exchange} \propto \pm\int U_{p-e}2\phi_1\phi_2\,dr \tag{3.83}$$

If the molecular orbital is symmetric, this term has the same sign as that upon which the Coulombic interaction between the electron and the protons depends and adds to it: if it is anti-symmetric it has the opposite sign and opposes it.

$$\begin{aligned}\left(U_{p-e}\right)_{Symmetric} &= \left(U_{p-e}\right)_{Coulomb} + \left(U_{p-e}\right)_{Exchange} \\ \left(U_{p-e}\right)_{Anti-symmetric} &= \left(U_{p-e}\right)_{Coulomb} - \left(U_{p-e}\right)_{Exchange}\end{aligned} \tag{3.84}$$

Thus, the total potential energy of the electron-proton interactions in the H_2^+ molecule depends on whether the molecular orbital is symmetric or anti-symmetric.

This has the effect of splitting the molecular ground state into two energy levels.

Since the Coulombic interaction is attractive, its potential energy will be negative: $\left(U_{p-e}\right)_{Coulomb} < 0.$.

Likewise, the exchange energy: $\left(U_{p-e}\right)_{Exchange} < 0.$

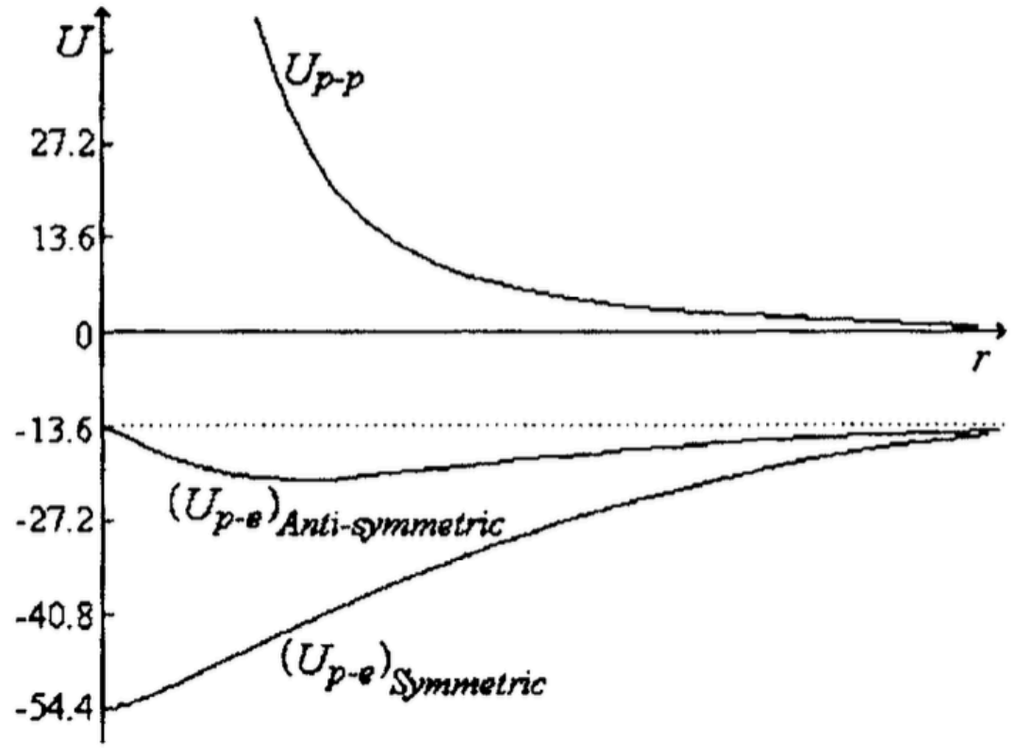
Hence, the effect of the minus sign in the expression for the potential energy of the electron-proton interactions of the anti-symmetric orbital -equation (3.84)-is to raise its value above that of the symmetric orbital:

$$\left(U_{p-e}\right)_{Anti-symmetric} \geq \left(U_{p-e}\right)_{Symmetric} \tag{3.85}$$

Thus, the higher energy level corresponds to the anti-symmetric molecular orbital and the lower energy level to the symmetric orbital.

The closer the nuclei are to one another, the greater the contribution of the exchange energy and the wider the gap between these energy levels (Fig 3.38).

Fig 3.38 The potential energy of the three basic interactions in the H_2^+ molecule as a function of the intemuclear separation, r. $U_{p\text{-}p}$ is the energy of the Coulombic interaction between the positive nuclei, $(U_{p-e})_{Symmetric}$ that of the proton-electron interaction and $(U_{p-e})_{Anti-symmetric}$ that of the anti-symmetric proton-electron interaction. When the intemuclear separation is large, $r \rightarrow \infty$, the energy of the proton-electron interaction is essentially just that of the 1s ground sate of the hydrogen atom, i.e., -13.6eV



The total potential energy of the anti-symmetric orbital in the H_2^+ molecule is

$$U_{Anti-symmetric} = (U_{p-e})_{Anti-symmetric} + (U_{p-p})_{Coulomb} \tag{3.86}$$

This function does not exhibit a minimum at any point.

Hence, it cannot be the function that describes the formation of a bond between the nuclei.

For this reason, the anti-symmetric orbital is termed an *anti-bonding orbital*.

The total potential energy of the symmetric molecular orbital is

$$U_{Symmetric} = (U_{p-e})_{Symmetric} + (U_{p-p})_{Coulomb} \tag{3.87}$$

This function exhibits a distinct minimum at U = -16.25eV and r = 0.106nm; the bond energy is 2.65eV (Fig 3.39).

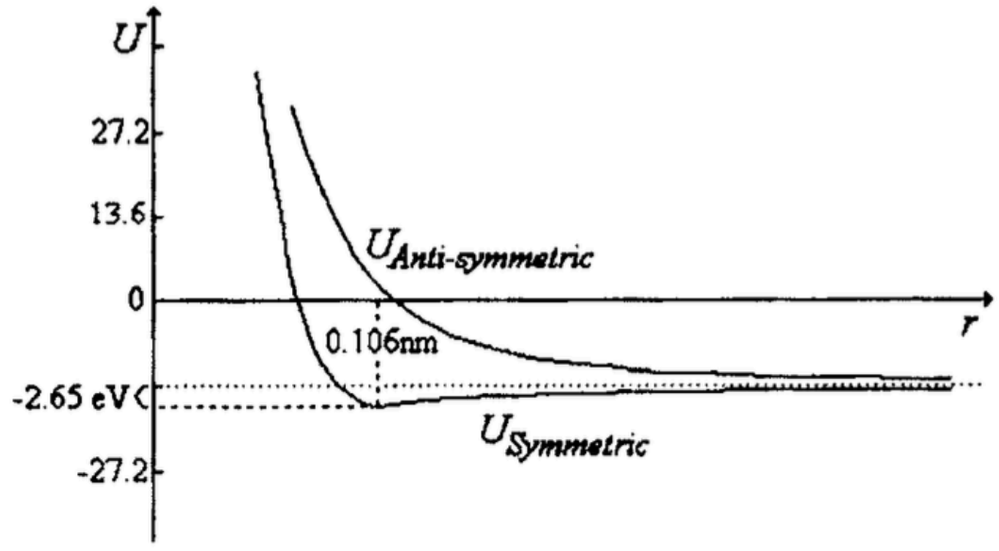


Fig 3.39 The potential energy diagram of the symmetric (bonding) and anti-symmetric (anti-bonding) orbitals in the H_2^+ molecule.

It is the function that describes the formation of the bond in the H_2^+ molecule. Hence, the symmetric orbital is termed a *bonding orbital*.

The bond formed in the H_2^+ molecule is exceptional in that it is a one-electron bond.

The simplest case of a typical covalent two electron bond is that found in the hydrogen molecule, H_2.

Because of the presence of the second electron in this molecule, three more factors must be taken into account when considering its nature:

1. **A second Coulombic repulsive interaction between the electrons. This acts to raise the system's total energy.**
2. **More base states that must be taken into account when considering the molecular wavefunction. This increases the number of interference terms, which, in turn, enhances the contribution of the exchange energy to the system's total energy. In the hydrogen molecule this *resonance* between the base states provides some 90% of the of the bond's total energy of 4.75eV; Coulombic effects contribute just 10%.**
3. **The overall wavefunction that describes the electrons comprising the bond must be anti-symmetric.**

The repulsive interaction between the electrons adds a third term to the general expression for the system's potential energy:

$$U = U_{p-p} + U_{p-e} + U_{e-e} \tag{3.88}$$

Substituting this expression in the appropriate Schrodinger equation produces an equation for which no exact solution has been found.

However, by restricting the number of base states taken into account, a solution is obtained that is qualitatively similar to that found for the H_2^+ molecule.

Above all, the inference that it is the symmetric orbital, ϕ_S, that leads to a bonded state remains valid.

Thus we see that covalency is predominantly a quantum effect, a consequence of the fact that the electrons comprising the covalent bond are exchangeable.

Though each originally belonged to one or other of the atoms, they lose their individual identities when the bond is formed.

Since the overall wave function that describes the pair of electrons must be anti-symmetric, the total spin function must be anti-symmetric; the electrons must be in a singlet state.

Thus, the quantum state that leads to bonding is that in which the electronic spins are opposed (Fig 3.40).

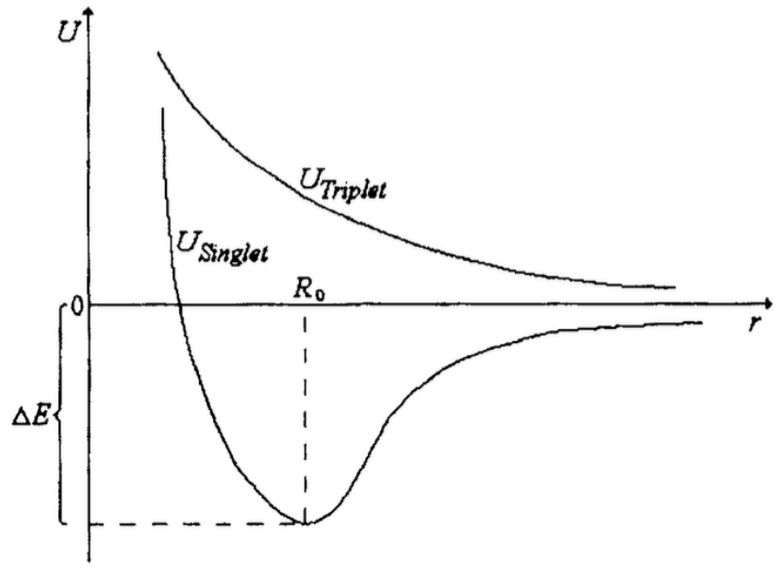


Fig 3.40 The potential energy diagram of a pair of hydrogen atoms. Whereas the energy of the triplet state rises monotonically, that of the singlet state exhibits a minimum at an internuclear separation of R_0 = 0.074nm. The corresponding bond energy ΔE = 5.3eV.

Like the bonds that bind its atoms, the actual structure of a molecule can also be regarded as a hybrid, formed by the superposition of all the conceivable physically indistinguishable structures the molecule can have.

This is the principle that underlies the concept of resonance postulated by Pauling for elucidating chemical structures.

In order that they be physically indistinguishable, the contributing structures can only differ in the positions of their electrons not in those of their atoms.

Likewise, each must have the same number of paired and unpaired electrons as otherwise they would be magnetically distinguishable.

The resultant spatial wave function is obtained by the weighted linear superposition of the spatial wave functions of the contributing structures.

In determining which of the possible wave functions correctly describes a particular system, the principle adopted is:

> ***The actual structure of the normal state is that one, of all conceivable structures, that gives the system the lowest energy and the maximum stability.***

The classical example of this technique is given by its application to the structure of the benzene molecule.

The conventional Kekule structural formula shows six quadrivalent carbon atom linked in a ring by alternating single and double bonds (Fig 3.41).

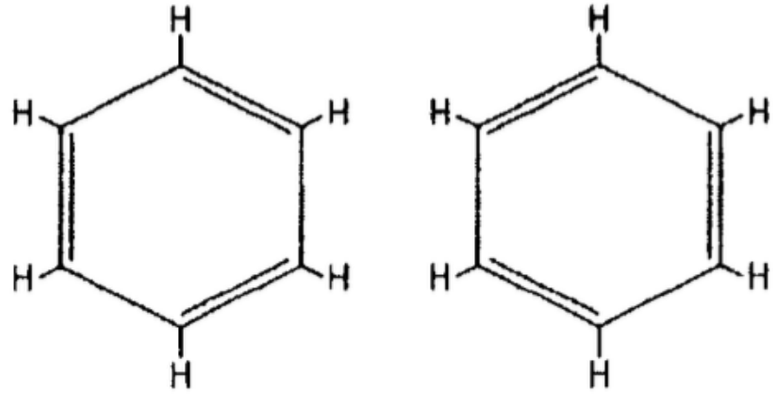


Fig 3.41 The two possible forms of the Kekule formula for benzene.

However, the chemical properties of benzene belie such distinct and localized bond-types.

Moreover, the actual energy of a benzene molecule-its *resonance energy*-is about 1.6eV less than it should be were this its structure.

According to Pauling's concept of resonance, the actual structure of the benzene molecule is that represented by the weighted linear superposition of the wave functions of its base states.

Two possible base states are the two forms of the Kekule formula.

These structures possess the same energy and so make equal contributions to the final state.

Like the other two-state systems we have examined, such as the H_2^+ molecule, the linear superposition of their respective wave functions gives rise to two distinct stationary states, one with an energy above the expected value and one with an energy below it.

The lower state can be described by the orbital $\phi_I + \phi_{II}$; the higher state by $\phi_I - \phi_{II}$.

In the case of benzene the difference between the energy levels is about 4.7eV. At normal temperatures, the benzene molecule is in the lower energy stationary state.

The existence of the higher energy level is evidenced by the fact that benzene absorbs ultra-violet radiation of frequency 1.15×10^{15}Hz.

In the lower stationary state, the six electrons comprising the double-bonds - the π electrons- are free to circulate around the ring, above and below the plane of the ring.

This delocalization effectively enlarges the potential well in which these electrons can move and accounts for the reduction in the molecule's energy.

Part 4 Interactions of Electromagnetic Radiation and Matter

In Nature, energy is being continuously exchanged, between the atoms and molecules that comprise material objects and the electromagnetic radiation that fills the space around them.

The interaction of sunlight with the objects on the earth's surface accounts for most of the phenomena we see around us, including life itself.

It was in order to account for the interaction between heat radiation and matter that the quantum hypothesis was first put forward by Planck.

Subsequently, Einstein extended Planck's idea to account for the photoelectric effect.

But it was only when Bohr combined the quantum hypothesis with the nuclear model of the atom in order to explain atomic spectra and to provide the first consistent atomic theory, that its full power as an investigative and interpretive tool became apparent.

In the following discussion we will show how this synthesis has been applied to the elucidation of a number of physical phenomena, all of which involve an interaction between radiation and matter.

We will restrict ourselves, at this stage, to phenomena involving radiation in the microwave, infra-red, visible and ultra-violet regions of the electromagnetic spectrum.

Interactions between matter and high energy radiation such as X-rays and y-rays are discussed in Part 5 of this discussion.

Many of the research techniques employed in the earth and life sciences, such as spectroscopy and fluorescence labelling, are based on these phenomena.

These too are dealt with in the following discussion.

In the second half of this century, the synthesis of quantum theory and the nuclear atom was widened to encompass the theory of relativity.

The new quantum theory of light and matter that this spawned is known as Quantum Electrodynamics or QED for short.

This "strange theory of light and matter", as one of its creators dubbed it, "describes all the phenomena of the physical world except the gravitational effect...and radioactivity." In homage to this remarkable union, a section is devoted to a brief account of its principles.

Part 4.1 The Passage of Radiation through Matter

In general, the passage of a beam of electromagnetic radiation through a material medium attenuates its intensity (Fig 4.1).

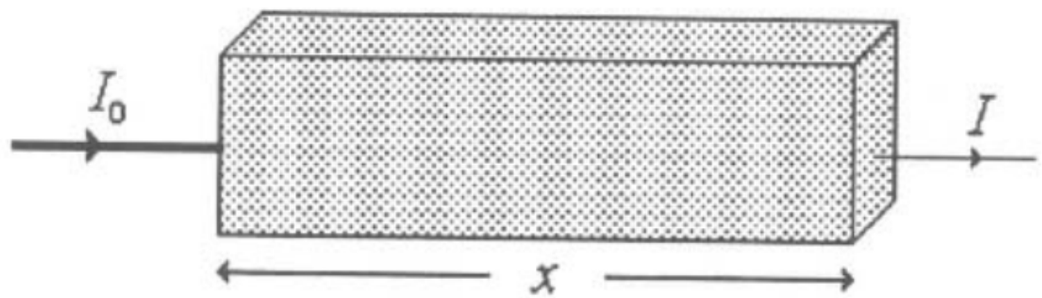


Fig 4.1 The passage of a beam of radiation through matter. After passing through a thickness x, the intensity, I_0, of the incident beam is reduced to I.

The reduction in the beam's intensity depends on the following factors:

(i) the extent of the beam's passage through the material medium;

(ii) the nature of the material medium;

(iii) the frequency of the radiation.

Absorption is the general term used to describe this reduction in intensity.

The absorption of radiation by matter is an example of *energy resonance*.

Atoms and molecules absorb radiation preferentially when its frequency coincides with one of their natural frequencies.

Selective absorption is the preferred absorption of radiation of a particular frequency.

Brick walls absorb visible light but not radio waves and so radio transmissions can be received inside a building even without an external aerial.

Glass is transparent to visible light but not to infra-red radiation; it does not absorb visible light but it does absorb infra-red radiation.

This accounts for the warming of glass-houses and of cars left out in the sun-the glasshouse effect.

Sunlight enters through their windows and is absorbed by their contents.

These warm up and, in turn, emit infra-red radiation, but this radiation cannot escape back through the glass.

It is trapped and raises the inside temperature still more.

White light, such as sunlight, is a mixture of colored lights-the colors of the rainbow.

It comprises a continuous visible spectrum that extends from the color red through orange, yellow, green, blue and violet.

Objects that are not themselves light sources are only visible by virtue of the light they reflect or scatter.

An object that returns all the light incident upon it is a white object; one that returns none of the light is a black object.

Colored objects absorb light of all colors except their own, i.e., they selectively absorb all the colors in sunlight except those they return to the eye.

For example, a green leaf selectively absorbs the red and blue constituents of white sunlight; the green constituent is left as the dominant color in the light it returns and so the leaf appears green to our eyes.

Substances which give a color to objects are called *pigments*.

Pigment particles selectively absorb certain of the constituents of white light and transmit the rest; the pigment's color is that of the transmitted constituents.

Some of the white light incident on a colored object or particle is reflected at its surface; the remainder enters the surface and its unabsorbed constituents are subsequently returned.

Thus, when a blue copper sulphate crystal is ground up a white powder is obtained.

Incident white light cannot penetrate into the powder's fine crystals and is scattered (reflected) in all directions at their surfaces.

Consequently, little or no light enters the crystals or is absorbed by them.

The color of a green object such as a leaf is called *body color*.

A thin gold foil appears yellow when looked at by the light it reflects; this is its surface color.

However, the foil appears green when it is used as a light filter and a source of white light is viewed through it; this is its body color.

Gold foil reflects yellow light but transmits green light.

The absorption of radiation by matter is typically studied by means of a *spectrophotorneter*.

The arrangement of a simple spectrophotometer that operates in the visible and near ultra-violet range is shown schematically in Fig 4.2.

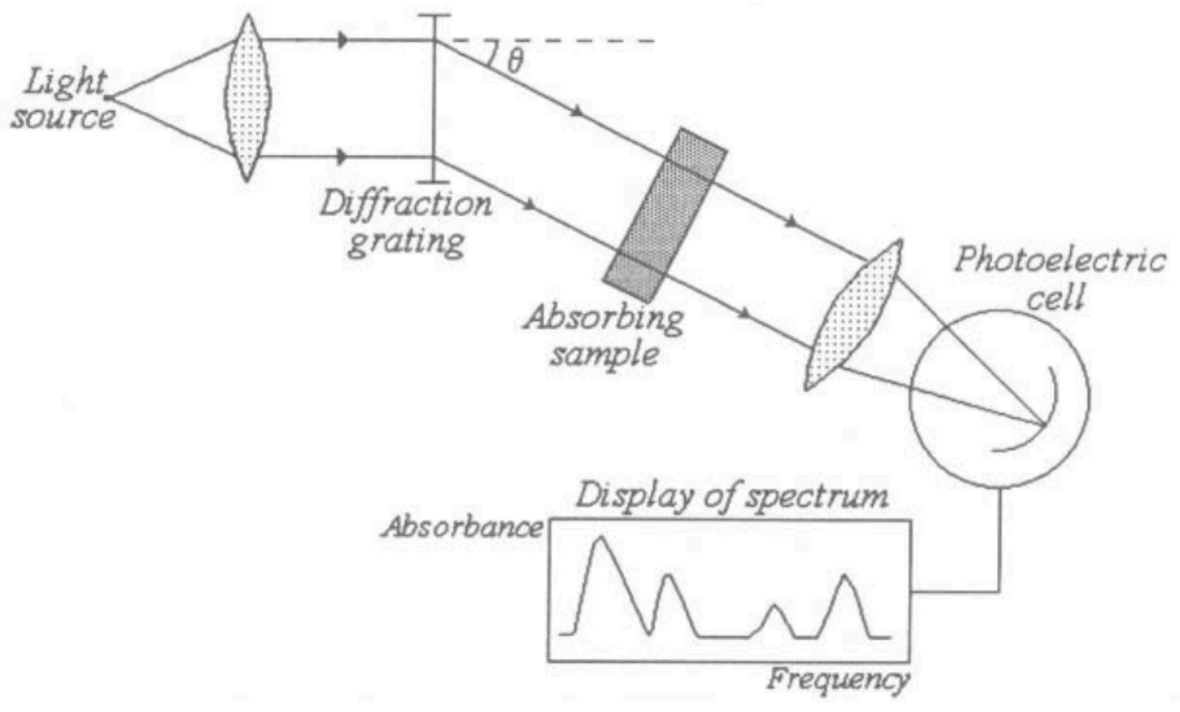
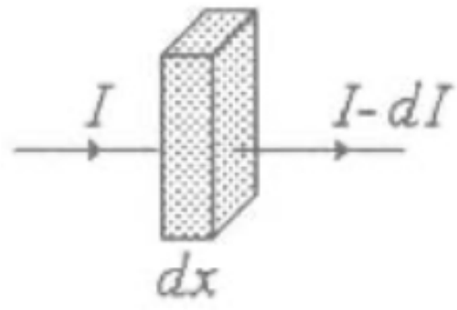


Fig 4.2 A diagram of a spectrophotometer that operates in the visible region. The assembly of the diffraction grating and/or the absorbing sample, together with the photoelectric cell, is rotated so that the sample is successively exposed to different parts of the spectrum. The intensity of the light emerging from the sample is measured by the photoelectric cell. The absorption spectrum is usually displayed in terms of the % absorbance as a function of the radiation's frequency.

4.1.1 The Attenuation of Radiation by Matter

The more absorbers there are in the radiation's path the greater is the probability it will be absorbed.



Thus, irrespective of the mechanism of its absorption, the reduction, -dI, in the intensity of a beam of radiation as it passes through an infinitesimal element of matter of thickness dx is given by

$$-dI = \mu I dx \tag{4.1}$$

where I is the intensity of the incident beam.

The constant of proportionality μ is called the *absorption* or *attenuation coefficient*; it has the SI units m^{-1} but is usually given in cm^{-1}.

The solution of equation (4.1) is obtained from the integration:

$$\int \frac{dI}{I} = -\mu \int_0^x dx \tag{4.2}$$

which gives:

$$I(x) = I_0 e^{-\mu x} \tag{4.3}$$

where I_0 is the beam's initial intensity and I(x) is its intensity after passing a distance x through the matter (Fig 4.3)

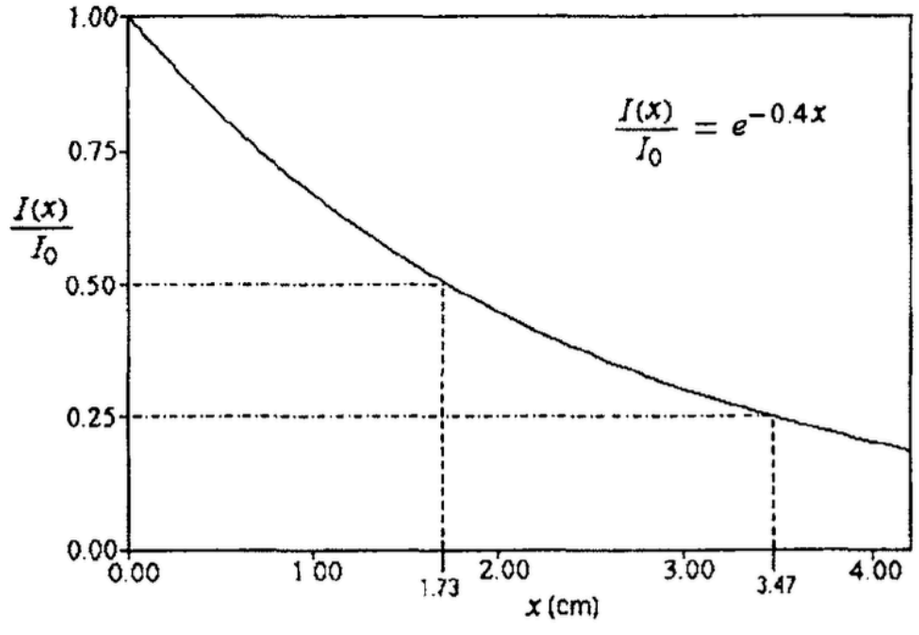


Fig 4.3 The attenuation of the intensity, I, of a beam of radiation as a function of the thickness, x, of the material medium in the case where attenuation coefficient $\mu = 0.4\text{cm-}1$.

The quotient

$$\log_{10}\left(\frac{I_0}{I(x)}\right) = \frac{\mu x}{2.303} \tag{4.4}$$

is called the *absorbance* or *optical density*, (O.D.).

In dilute solutions, where the solute particles are the principle absorbers, the absorption coefficient, μ, depends on their molar concentration, C:

$$\mu = 2.303\varepsilon C \tag{4.5}$$

The quantity ε is called the *molar extinction coefficient*. Combining equations (4.4) and (4.5) gives:

$$I(x) = I_0 e^{-2.303\varepsilon Cx} \tag{4.6}$$

In this form, the relationship is called *Beer-Lambert 's Law*.

4.1.2 Mechanisms of the Absorption of Radiation

Radiation in the microwave, infra-red, visible and ultra-violet regions of the electromagnetic spectrum is absorbed by one or more of the following three mechanisms: ionization, excitation and scattering.

Each of these mechanism is characterized by its own particular absorption coefficient.

An absorbing medium's overall absorption coefficient, μ, is the sum of these coefficients:

$$\mu = \mu_{Scattering} + \mu_{Excitation} + \mu_{Photoelectric\ Effect} \tag{4.7}$$

1. Ionisation

This can occur if the energy of the photons in the incident radiation is equal to or greater than the ionisation energy of the atoms or molecules in the absorbing medium. Electrons are then released from the atoms and molecules and photons are annihilated, resulting in a corresponding reduction in the intensity of the beam of radiation.

2. Excitation

At room temperature, the atoms and molecules that comprise matter are usually in their ground state. Photons possessing the right amounts of energy, can excite the atoms and molecules to higher energy levels. When this occurs, radiation is absorbed and the intensity of the transmitted beam is reduced accordingly.

This mechanism is very selective; a given material only absorbs certain frequencies. Consequently, this phenomenon can be made the basis for the identification of materials and for determining the concentration of the absorbing solutes in solutions.

Excitation occurs in gases when the energy of the photons in the incident radiation exactly equals the difference between the energy of an excited state and that of the ground state. Only such photons are absorbed by the gas and so the absorption spectra of gases are line spectra that comprise discrete frequencies.

Because of the much more complicated patterns of their energy levels, liquids and solids absorb whole ranges of frequencies. Thus, their absorption spectra are typically *band spectra* (Fig 4.4). Nevertheless, the absorption is still sufficiently selective in many cases for these spectra to still be used for identifying the materials.

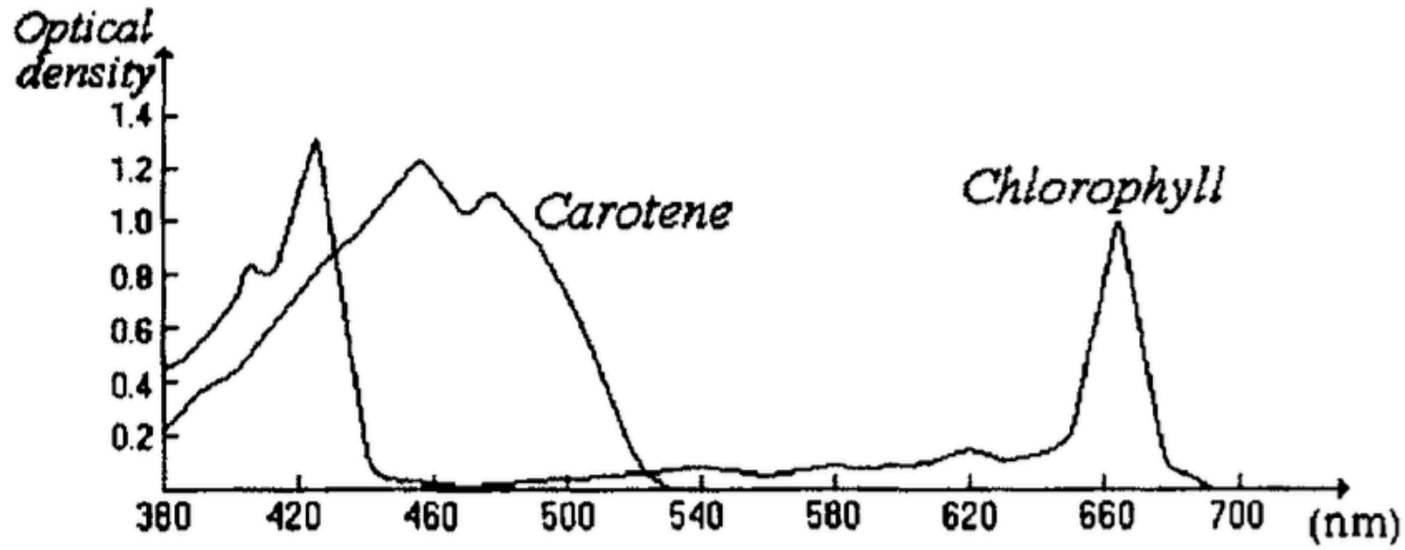


Fig 4.4 The absorption spectrum of chlorophyll and carotene in the visible range. Chlorophyll has absorbance maxima at around 435nm and 665nm. Carotene has absorbance maxima at around 450nm and 490nm.

Having risen to an excited state, the atom or molecule can lose its excess energy and return to the ground state in three different ways:

(i) By non-radiative transitions such as those that result from collisions between the excited particles and other atoms or molecules in the material. In the process, the excess energy is transferred and shared out between other energy levels in the system. The overall result is an increase in the average kinetic energy of the particles comprising the matter, i.e., the conversion of the incident radiation's energy to heat energy. This is the most common method by which excited particles lose their excess energy.

(ii) The excess energy is converted to chemical energy. For example, the excess energy possessed by the excited particle can initiate a chemical reaction such as a photochemical reaction.

(iii) The excited particles may return to the ground state by spontaneous or stimulated emission of a photon. Generally, the photon will be emitted in a different direction from that of the incident beam.[4]

The average lifetime of an excited state is from 10^{-9} to 10^{-6} seconds. If nothing disturbs the particle during this time, it ultimately returns to the ground state by emitting a photon, namely, by the third of the above methods. However, this does not usually happen. During its lifetime, short though it may be, thousands of collisions will have taken place between the excited particle and the other particles in the system. Consequently, it is by the first method—non-radiative transitions—that excited particles most commonly lose their excess energy.

3. Scattering

We discussed the scattering of electromagnetic radiation by free electrons in Chapter 2.3 under the heading of the Compton Effect. Here we will be concerned with the scattering of radiation by bound electrons. This phenomenon is most easily interpreted in terms of the wave theory of radiation.

When a periodic electromagnetic wave impinges on an object, the electric and magnetic fields associated with it interact with the electrons in the atoms it comprises. To a first approximation, the effect of the magnetic field can be ignored, its magnitude in the case of a plane electromagnetic wave being of the order $1/c$ that of the electric field. The electric field induces oscillating electric dipoles in the object's atoms and molecules. Acting like antennas, these dipoles in turn radiate secondary or *scattered* waves in all directions. As a result, the intensity of the radiation in the direction of the original beam is attenuated. The frequency of the scattered waves is generally the same as that of the original wave.

The dipole oscillations are induced parallel to the electric field of the incident wave, i.e., in the plane normal to its direction of propagation. Consequently, the secondary radiation is emitted at maximum intensity in the direction perpendicular to the dipole's axis and at minimum intensity in the direction parallel to the axis. Because of this, the scattered radiation is always polarised to some degree (Fig 4.5).

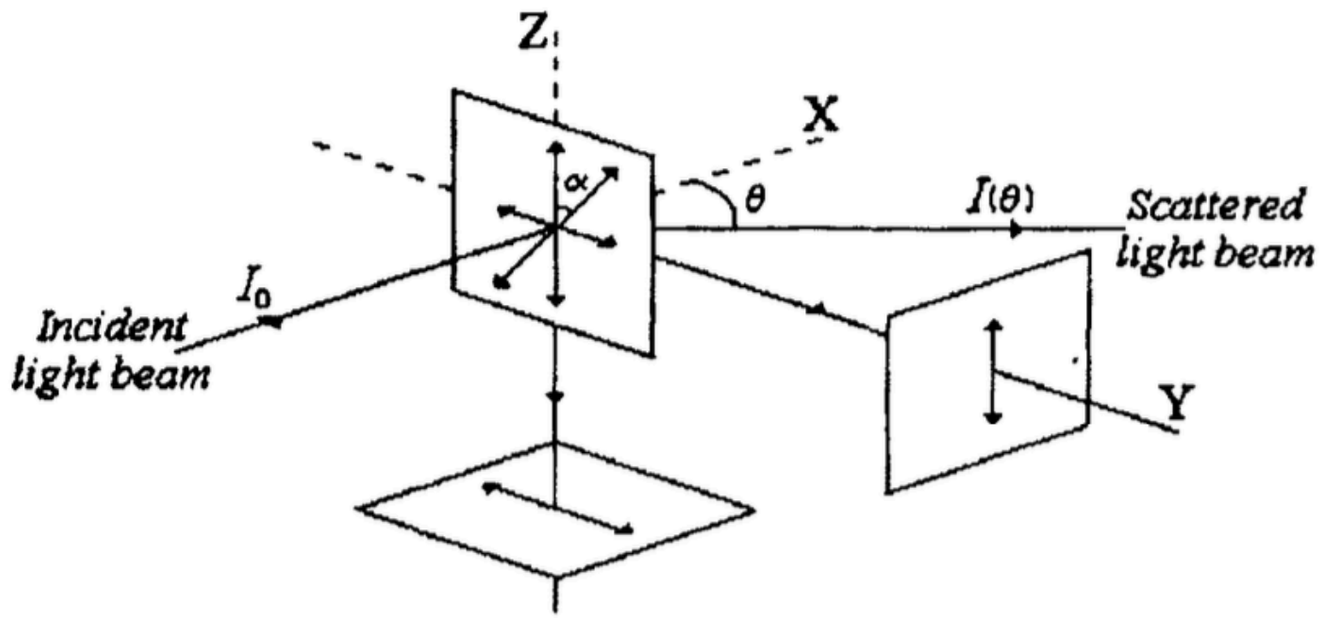


Fig 4.5 The scattering of a beam of unpolarized electromagnetic radiation. The incident beam of intensity I_0, propagates along the X axis. The figure shows an oscillating electric dipole that is induced in the ZY plane at an angle α to the 2 axis. The linearly polarized radiation scattered along the Y axis is that produced by the z component of the dipole. Similarly, the linearly polarized radiation scattered along the Z axis is that produced by the y-component of the dipole. The radiation that continues along the X axis is unpolarized. The radiation of intensity $I(\theta)$ scattered at an angle θ is partially polarized.

The intensity of the scattered radiation is proportional to the square of the amplitude of the vibrations of the induced electric dipole.

The closer the frequency of the incident radiation to the natural frequency of the dipole, the greater the amplitude of the induced vibrations; this is a straightforward example of *amplitude resonance*.

The natural frequencies of atomic electric dipoles are in the ultra-violet region of the spectrum.

Hence, the blue and violet constituents in white light are scattered more than the shorter frequency red and yellow constituents.

This explains why when a beam of white light is scattered by smoke particles, the light scattered at right angles to the direction of the incident beam has a blue hue and is linearly polarized, whereas that which is transmitted along the original direction has a red/yellow tint and is unpolarized (Fig 4.6)

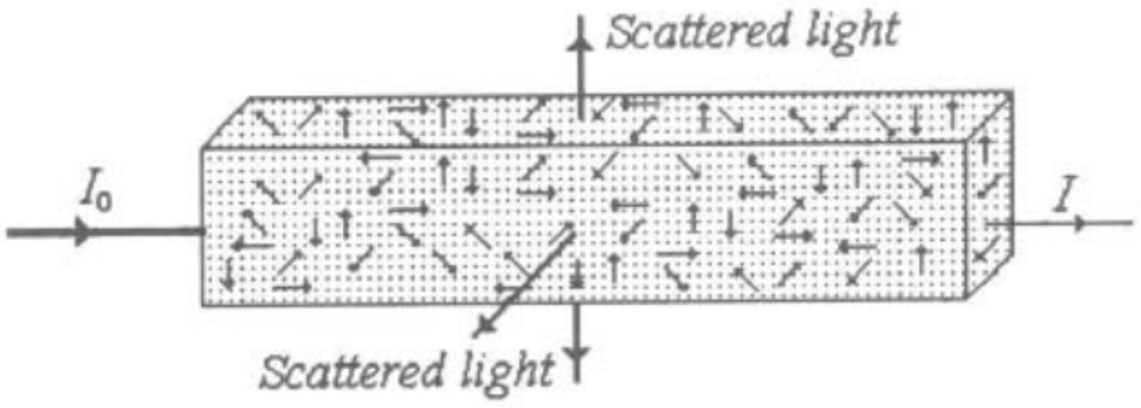


Fig 4.6 The scattering of an incident beam of white light of intensity I_0 by smoke particles. The scattered light of overall intensity I_S is polarized and has a blue hue. The transmitted beam of intensity $I = I_0 - I_S$ is unpolarized and has a red/yellow tint. The intensity of the transmitted beam decreases exponentially with the length of its path through the smoke.

When viewed from an orbiting spacecraft or from the moon's surface, the sky -the dome overhead- is black at day as well as at night.

The blue of the sky seen from the earth's surface is a result of the scattering of sunlight by the atmosphere.

The blue and violet constituents of sunlight, i.e., the shorter wavelengths, are scattered more by the particles comprising the earth's atmosphere than are the other constituents.

Thus, an observer on the earth's surface sees a blue sky overhead in all directions, except in that directly towards the sun.

At sunrise and sunset, when the sunlight's path through the atmosphere to the observer's eye is longest, even more of the blue and violet constituents are scattered, leaving the transmitted beam with little of these colors.

Thus, at these times, the sun's disc appears red.

The blue of the sky is also polarized, the direction of polarization at each point changing during the day as the sun moves across the sky.

It has been suggested that this changing pattern in the sky's polarization is used as a navigational reference grid by those creatures, such as bees, whose eyes, unlike the human eye, can distinguish between different modes of light polarization.

With an absolutely homogeneous medium, such as perfect crystal, the secondary waves emitted by the oscillations induced in the electrons all interfere with one another destructively.

Hence, scattering is not observed.

The incident beam is just reflected and refracted according to the laws of geometrical optics. If the earth‘s atmosphere was absolutely homogeneous, the sky would also appear black to observers looking upwards from its surface.

The quantitative treatment of the scattering of light by particles, whose dimensions are small relative to the wavelength of the radiation incident upon them, was accomplished by Raleigh.

He showed that the intensity of the scattered light depends on $1/\lambda^4$ where λ is the wavelength of the incident light.

If the particles are larger, such as dust particles, Rayleigh’s treatment is no longer valid.

Under these circumstances, scattered white light is untinted; it is white.

Thus, though water vapor in the atmosphere contributes to the light scattering that produces the blue of the sky, when the vapor condenses into fine drops whose dimensions are relatively large, white clouds are seen.

Experiments show, that though all electromagnetic waves propagate through empty space with velocity c, their velocity, v, through material mediums depends on the nature of the material and on the frequency, *f*, of the wave; the phenomenon is called *dispersion*.

The ratio, n, between the two velocities is called the *refractive index*:

$$n = c/v \tag{4.8}$$

Refractive indices are always equal to or greater than unity.

The color of light, i.e., the energy of its photons and hence its frequency, is unaffected by the changes in velocity that occur when light passes from one medium to another; a blue swimming costume remains blue in the water.

However, if its wavelength is measured it is found to be different.

The light's wavelength, λ', in the material medium is always less than that its wavelength, λ, in empty space, such that the fundamental relationships

$$f = c/\lambda, \quad f = v/\lambda', \quad \text{and} \quad \lambda = n\lambda' \tag{4.9}$$

are respectively satisfied.

Dispersion arises from the fact that the secondary waves radiated by the vibrations of the induced dipoles in the material medium are out of phase with the incident wave; a forced oscillator is not generally in phase with the driving force.

As a result, the actual wave produced in the material medium from the superposition of the incident wave and the secondary waves exhibits a phase velocity, v, different from the phase velocity, c, of electromagnetic waves *in vacuo*.

Although no particle or signal can move at a velocity greater than that of electromagnetic waves *in vacuo*, c, the velocity of high energy charged particles in a material medium can exceed the phase velocity, v = c/n at which electromagnetic waves propagate in the medium.

Consequently, if the particle itself emits electromagnetic waves, these will form a shock wave analogous to that formed when an object moves at supersonic speeds through the air.

The electromagnetic shock waves, called *Cerenkov radiation*, propagate along conical surfaces that make an angle α relative to the direction of the particle such that

$$\sin\alpha = \frac{v_{radiation}}{v_{particle}} \tag{4.10}$$

This phenomenon has been employed in techniques for measuring the velocities of high energy elementary particles.

4.1.3 Quantum Electrodynamics

Interactions between light and matter, such as the dispersion and scattering of light, can also be elucidated in terms of a combination of the theory of relativity and quantum mechanics.

The procedure by which this is done is called *Quantum Electrodynamics*.

The principal assumption underlying this procedure is that all the interactions that occur between radiation and matter can be reduced to just three basic processes:

1. **A photon moves from one position in spacetime to another.**
2. **A electron moves from one position in spacetime to another.**
3. **An electron emits or absorbs a photon.**

These three basic processes are most conveniently depicted by Feynman diagrams as shown in Fig 4.7

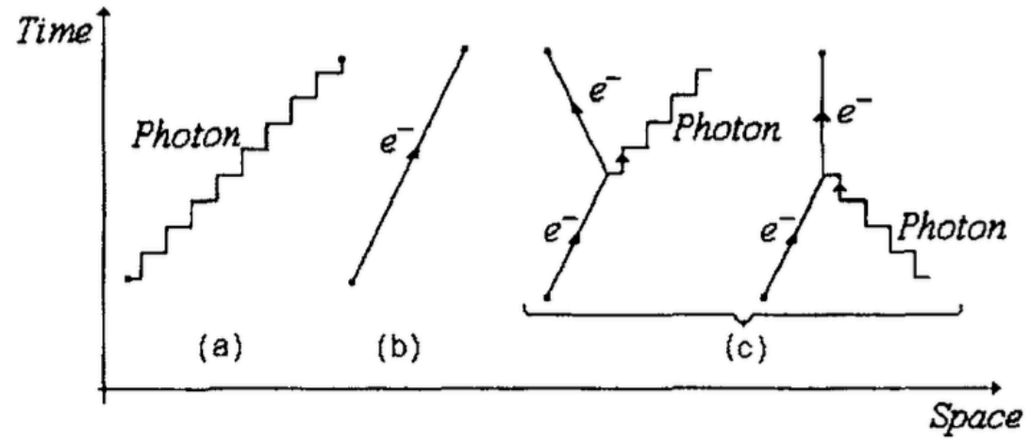


Fig 4.7 A Feynman (spacetime) diagram showing the three basic processes of quantum electrodynamics. (a) A photon moving from one point in spacetime to another. (b) An electron moving from one point in spacetime to another. (c) The emission or absorption of a photon by an electron.

A second assumption is that a probability amplitude (wave function) can be assigned to each of these basic processes and that the appropriate combination of these amplitudes gives the probability amplitude, Ψ, of a phenomenon or event.

For example, by making the appropriate combination of the probability amplitudes of the three basic processes, the overall probability amplitude for the passage of a photon through a dispersive medium from one point in spacetime to another or for the displacement of an electron in spacetime can be obtained.

In computing the overall probability amplitude, Ψ, all the combinations of these three processes that can give rise to the phenomenon or event are first specified and their separate probability amplitudes are calculated.

These are then combined to give the overall probability amplitude for the phenomenon or event.

This may all sound a very laborious technique but it has produced the most accurate numerical predictions yet made by any physical theory.

To illustrate this technique, we consider first the displacement of a photon from one point in spacetime to another.

The trajectory of a photon through spacetime is its worldline. In a regular Minkowski diagram, this world-line has a slope of 45º, indicating that the *in vacuo* velocity of the photon (light) is equal to c.

Over large distances this is quite correct and satisfactory.

However, assigning a trajectory to a photon's motion over the very short distances it travels between scattering events in a material medium is forbidden by quantum mechanics.

There is a non-zero probability of the photon appearing at any point in the medium after each scattering event.

Consequently, no definite conclusions can be drawn about the photon’s velocity between these events; it may be more, less or equal to c.

It follows, that in the Feynman (spacetime) diagrams that depict the photon-electron interactions that underlie the dispersion and scattering of light, the photon's world-line will not necessarily have a slope of 45° (Fig 4.8).

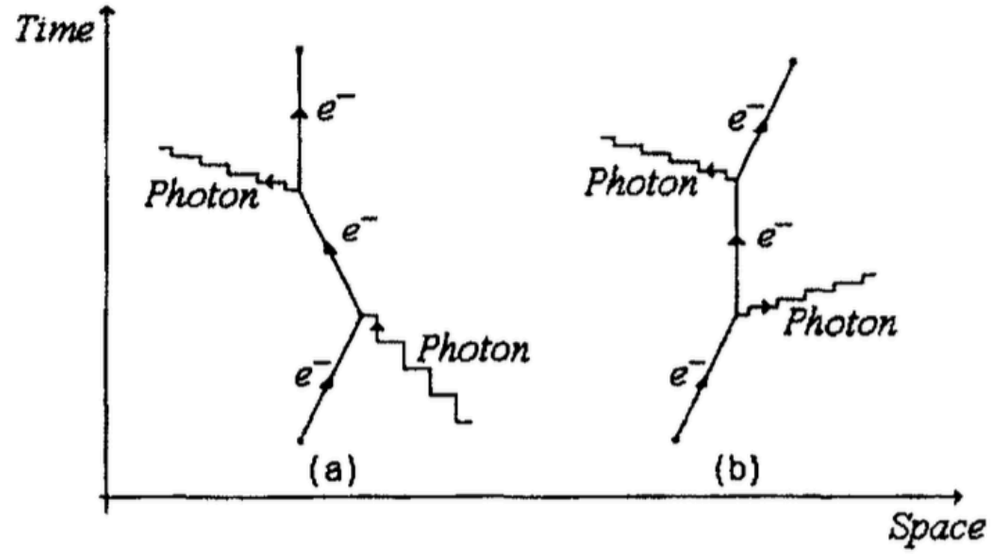


Fig 4.8 The scattering of light involves the absorption and emission of a photon by an electron not necessarily in that order. In (a) the photon is first absorbed and then emitted later. In (b) the electron first emits a photon and absorbs another identical one (though not necessarily the same one) later. Both of these alternative scattering mechanisms contribute to the overall probability amplitude for the phenomenon.

The probability amplitude, Π_m, for a photon known to be at point A in spacetime to appear at some other point B is an inverse function of the spacetime interval between the points.

The interval between points lying on a 45" world-line is zero; this corresponds to a velocity of c and it is this outcome that makes by far the greatest contribution to the probability amplitude n.

However, the small contributions made by other intervals, $I < 0$ and $I > 0$, that correspond respectively to velocities slower and faster than c, cannot be totally ignored, even though they generally cancel out over long distances.

Turning to the electrons, we recall that quantum mechanics forbids the assignment of a specific trajectory to the individual motions.

This is equivalent to saying that every conceivable trajectory is a possibility, though not all may be equally probable.

On this view, the displacement of an electron from one point in spacetime to another can be regarded as taking place in an infinite number of different ways: in a single direct hop or in any sequence of two or more hops (Fig 4.9).

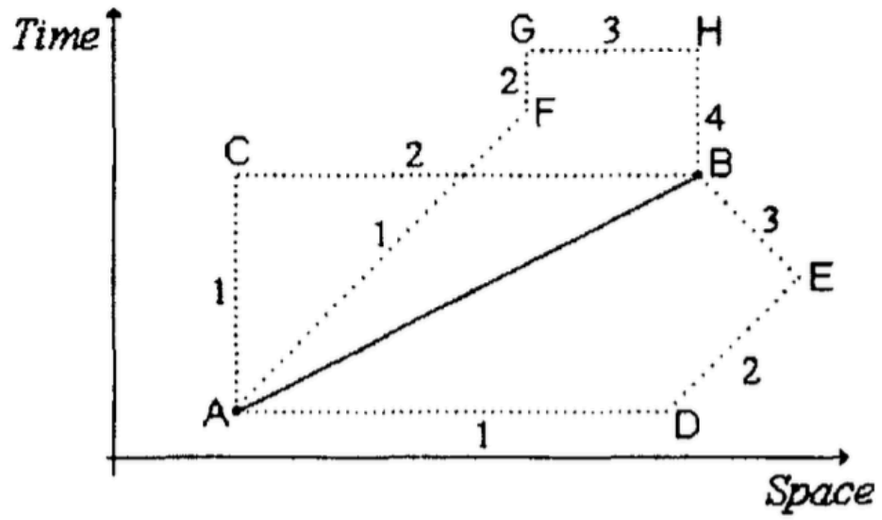


Fig 4.9 The displacement of a particle from point A to point B can be envisaged as taking place in a single hop or by an infinity of alternate paths each involving more than a single hop. Three such alternative paths are shown: ACB - two hops; ADEB - three hops, AFGHB four hops. Note that even hops that show the particle going backwards in time such as the HB hop must be taken into account.

The probability amplitude for each discrete hop is the same as that for a photon to go from one point in spacetime to another.

Thus, the probability amplitude for an electron known to be at point A in spacetime to appear at some other point B is Π_{AB}.

However, those paths that comprise more than a single hop will also involve changes in the electron's direction.

This too must be taken into account in calculating the probability amplitude for its displacement.

This is done by assigning a probability amplitude to the change in direction.

This probability amplitude is the square, m^2, of a number m whose value is associated with the particle's mass. In the case of a photon the value of m is zero.

In accordance with the rules for the combination of probability amplitudes (wave functions), the probability amplitude, Φ_{AB}, for the displacement of an electron from point A to point B will be given by the sum of a series of terms each of which is the probability amplitude for one of the possible paths the electron could take:

$$\begin{aligned}\Phi_{AB} \propto & \Pi_{AB} \\ & + \Pi_{AC} \cdot m^2 \cdot \Pi_{CB} \\ & + \Pi_{AD} \cdot m^2 \cdot \Pi_{DE} \cdot m^2 \cdot \Pi_{EB} \\ & + \Pi_{AF} \cdot m^2 \cdot \Pi_{FG} \cdot m^2 \cdot \Pi_{GH} \cdot m^2 \cdot \Pi_{HB} \\ & + \quad \cdots\end{aligned} \tag{4.11}$$

But this is not the end of the story.

Another factor that must be considered is the possibility that the electron may emit or absorb photons during its displacement.

This is the third of the basic processes postulated by quantum electrodynamics.

A probability amplitude is assigned to this process too.

The probability amplitude for the emission or absorption of a photon by an electron is a number, j, called the *coupling number*.

The numerical value o f j is about -0.1.

The photon emitted by one electron may be reabsorbed by the same electron or may be taken up by another electron.

The exchanged photon does not appear in either the initial or final states of the system and so is undetectable; such photons are usually referred to as *virtual photons*.

Limiting ourselves to the three alternative routes shown in Fig 4.10, we can now demonstrate how to compute the probability amplitude, Ψ , that describes the transformation of a system of two electrons at points A and C into a system of two electrons at points B and D.

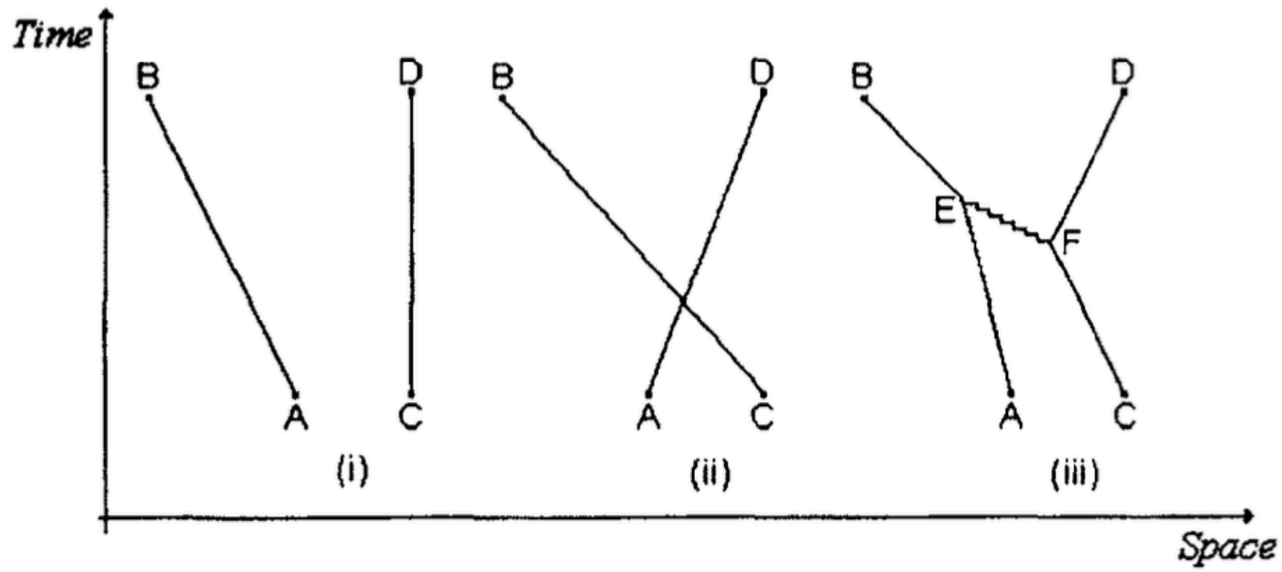


Fig 4.10 Three alternative routes by which a system of two electrons at points A and C can become a system of two electrons at points B and D.In the first the electrons move directly; one electron goes from A directly to B, the other from C directly to D. In the second they 'cross-over'. In the third a photon is exchanged between the electrons at some point during their passage from the initial to final points.

The first step is to specify the probability amplitude of each of the first two alternatives.

(i) The two sub-events that comprise the first alternative occur concomitantly, and so the probability amplitude, Ψ(i), for this way is the product of the probability amplitudes, Φ_{AB} and Φ_{CD}, of the two sub-events.

$$\Psi(\mathrm{i}) = \Phi_{AB} \cdot \Phi_{CD} \qquad (4.12)$$

(ii) Similarly, the two sub-events that comprise the second alternative occur concomitantly, and so the probability amplitude, Ψ(ii), for this event is the product of the probability amplitudes, Φ_{AD} and Φ_{CB}, of the two sub-events.

$$\Psi(\mathrm{ii}) = \Phi_{AD} \cdot \Phi_{CB} \qquad (4.13)$$

Combining these two alternatives gives a first fair approximation for the probability amplitude of the process:

$$\Psi = \Psi(\text{i}) - \Psi(\text{ii}) \tag{4.14}$$

The probability amplitudes (wave functions) are subtracted because the electrons are fermions and so the function Ψ must be anti-symmetrical.

(iii) The result we have obtained can be improved by taking into account the third alternative shown in Fig 4.10, namely, the exchange of a photon at some point in spacetime. The probability amplitude for this option is:

$$\Psi(\text{iii}) = \Phi_{AE} \cdot j \cdot \Phi_{EB} \cdot \Phi_{CF} \cdot j \cdot \Phi_{FD} \cdot \Pi_{EF} \tag{4.15}$$

where Φ_{AE} and Φ_{EB} are the probability amplitudes for the hops from A to E and from E to B respectively; likewise, Φ_{CF} and Φ_{FD} are the probability amplitudes for the hops from C to F and from F to D respectively; and Π_{EF} is the probability amplitude for the exchange of the virtual photon.

Including the third alternative gives for the probability amplitude of the process:

$$\Psi = \Psi(\text{i}) - \Psi(\text{ii}) + \Psi(\text{iii}) \tag{4.16}$$

By the use of powerful computers the technique can be continued and more alternatives can be taken into account, leading to better and better refinements in its predictions.

In practice, it is found that taking into account the entire infinity of ways in which a phenomenon or event can possibly occur produces meaningless results; such rigorous mathematical consistency removes any meaning the numbers m and j may possess.

This is really not surprising, for it implies that the interacting particles are point-sized; only a point-sized particle could change its direction or emit/absorb a photon at every conceivable mathematical point in spacetime.

But the magnitude of the electric field close to a point-sized electron would rise asymptotically to infinity.

This would mean that the electron has an infinite mass, which is obviously nonsensical.

This problem is avoided by limiting the total number of paths taken into account in the computations.

It transpires, that the values that must be assigned to m and j in order that the computations give the observed values for the mass and charge of the electron depend on the level at which the computations are ceased; different values are needed if all hops shorter than 10^{-30}m are disregarded than if the limit is set at say 10^{-40}m.

However, if the values of m and j obtained from the analysis of a particular phenomenon at a certain limit are used for the analysis of other phenomena down to the same limit, the new predictions also agree with experiment.

Thus it appears that the only factors that depend on the minimum length set for the hops are the values of the theoretical numbers m and j.

Everything else about the phenomena, and in particular their observables, seems to be unaffected by the limit chosen.

The numbers m and j are functions of just space and time.

Their effect is to change the phase of the particle's probability amplitude (wave function) at each change in direction or emission/absorption of a photon whilst leaving the probability density, $|\Psi|^2$, which is an observable quantity unchanged.

Theories in which all measurable quantities (observables) remain unchanged when the phases of the fields are altered by an amount that is a function of space and time are called gauge theories.

The observables are said to be phase symmetrical. Quantum electrodynamics is a gauge theory.

It is believed that gauge theories can provide the basis for a description of all elementary particle interactions.

The process by which the numbers m and j are determined is called *renormalization*.

It breaks down at extremely short distances, $\sim 10^{-75}$m, and high energies.

This breakdown cannot be cured by mathematical schemes, it is inherent in the structure of quantum electrodynamics.

A unified theory which allows for renormalization at all energy scales has yet to be formulated.

The greatest quantitative success of quantum electrodynamics has been its prediction of the gyromagnetic ratio of the electron.

This can be very accurately measured: the experimental value is γ = -2.0023193048(4).

The value predicted by quantum electrodynamics is γ = -2.0023193048(8) .

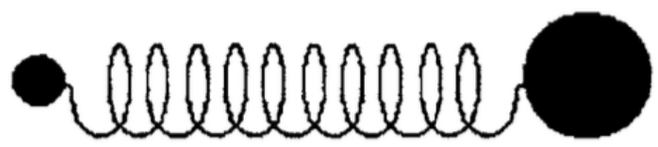
Because of the remarkably close correspondence between the predicted and experimental values, QED has been described as “the most accurate of all physical theories.”

Part 4.2 Molecular Spectra

4.2.1 Molecular Energies

We can picture a molecule as an assembly of small almost point masses-the nuclei of its component atoms, linked together by almost massless springs-the bonding electrons (Fig 4.11).

Fig 4.11 A diatomic molecule pictured as two spherical masses linked by a massless spring.



In this model, the total energy of a molecule, $E_{molecule}$, can be viewed as the sum of its electronic energy, $E_{electronic}$, its vibrational energy, $E_{vibration}$ and its rotational energy, $E_{rotation}$:

$$E_{molecule} = E_{electronic} + E_{vibration} + E_{rotation} \tag{4.17}$$

The vibrational and rotational energies are both related to the motion of the nuclei of the molecule's constituent atoms.

Like its electronic energy, a molecule's vibrational and rotational energies are also quantized, i.e., they can only take certain discrete values.

Thus, every molecule has a characteristic set of vibrational energy levels and rotational energy levels.

Typical values for the differences, ΔE_{elec}, ΔE_{vib} and ΔE_{rot}, between adjacent electronic, vibrational and rotational energy levels are, respectively:

$$\Delta E_{elec} = 1 - 10\text{eV}$$
$$\Delta E_{vib} \approx 10^{-1}\text{eV}$$
$$\Delta E_{rot} \approx 10^{-4}\text{eV}$$

i.e., $\Delta E_{elec} >> \Delta E_{vib} >> \Delta E_{rot}$.

Consequently, each electronic level can accommodate an array of vibrational sub-levels and, similarly, each vibrational level can accommodate an array of rotational sub-levels (Fig 4.12).

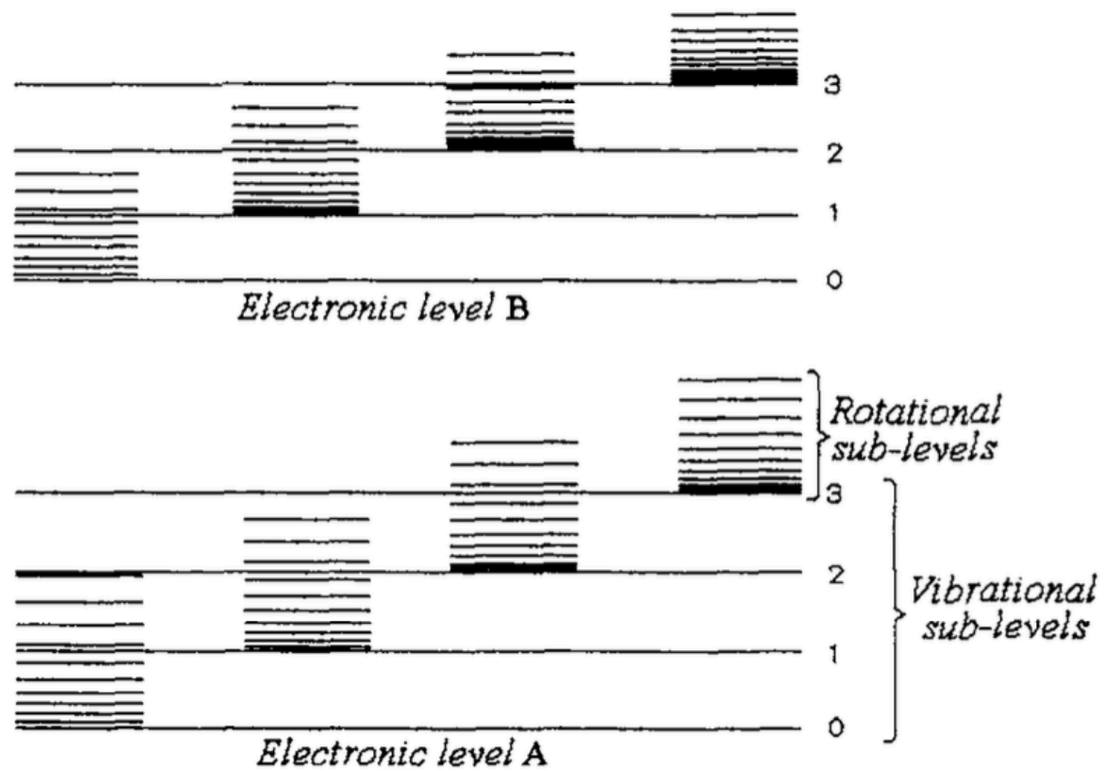


Fig 4.12 Two electronic energy levels, A and B, in a molecule. A number of vibrational energy levels, 1,2,3, ..., are associated with each electronic energy level. Similarly, a number of rotational sub-levels are associated with each vibrational energy level.

In common with atoms, molecules too can absorb and emit radiation by means of transitions between the energy levels associated with their electrons, i.e., through changes in their electronic energy.

However, there are two further modes by which molecules can absorb or emit radiation:

1. **By means of changes in the internal vibrational energy of the molecules.**
2. **By means of changes in the rotational energy of the molecules.**

The photons associated with transitions between electronic levels are in the visible and near ultra-violet range; those associated with transitions between vibrational levels in the same electronic level are in the infra-red; those associated with transitions between rotational levels in the same vibrational level are in the microwave region.

The probabilities of the various transitions are expressed by appropriate selection rules.

Transitions between a molecule's electronic energy levels are usually accompanied by changes in its vibrational and rotational energy.

Consequently, when a molecule 'falls' from an excited electronic state to a lower one, its vibrational and rotational sub-levels can also change.

Similarly, transitions between vibrational energy levels are usually accompanied by changes in the molecule's rotational energy.

Hence, transitions between vibrational levels often occur with changes in the rotational sub-level.

To a first approximation, only molecules possessing an electric dipole moment can absorb or emit electromagnetic radiation.

Consequently, non-polar diatomic molecules such as H_2 and Cl_2, and symmetric polyatomic molecules such as CO_2 and CH_4 do not exhibit rotational or vibrational spectra; radiative transitions between rotational and vibrational states in such molecules do not occur.

Non-radiative transitions can occur as a result of collisions.

Information about vibrational transfers in such molecules can be obtained from Raman scattering.

4.2.2 Rotational Spectra

Fig 4.13 shows a typical rotational absorption spectrum, that of gaseous hydrogen chloride, $HCl_{(g)}$. The spectrum exhibits two principle characteristics:

A. It comprises many lines, all of comparable intensity.

B. The absorption lines are equally spaced along the frequency axis.

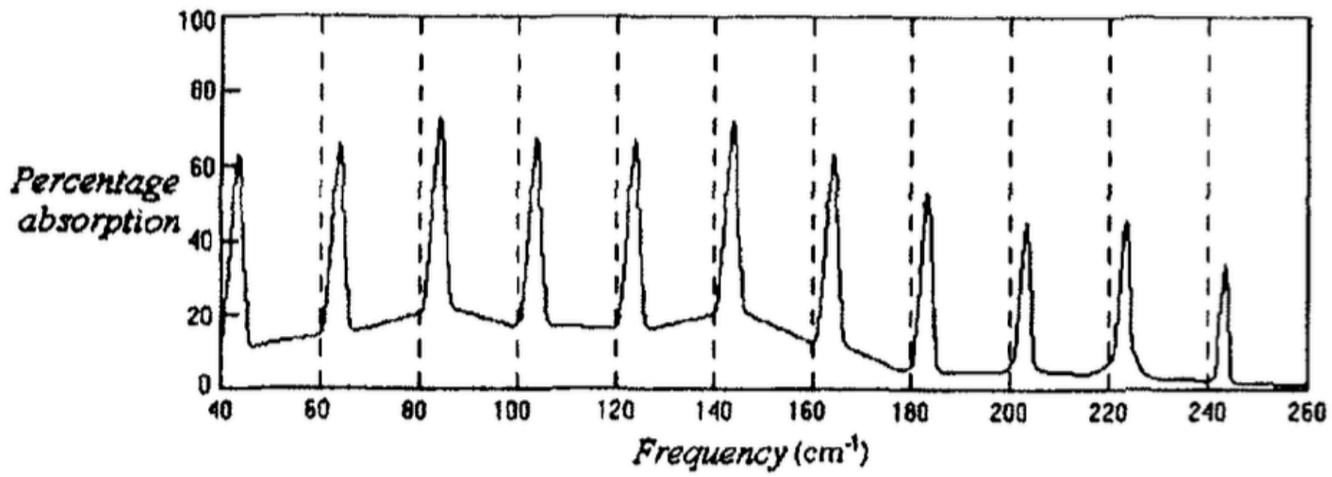


Fig 4.13 Rotational absorption spectrum of $HCl_{(g)}$.The horizontal frequency axis is scaled in units of the wave number,$1/\lambda$, where λ is the radiation's wavelength in centimeters ($1\ cm^{-1} = 3x10^{10}Hz$).

By analogy with the expression $K = p^2/2m$, that relates an object's kinetic energy, K, to its linear momentum, p , the rotational energy, E_{rot}, of a diatomic molecule ('dumb-bell'), such as that pictured schematically in Fig 4.14, revolving around an axis through its center of mass is given by:

$$E_{rot} = L^2/2I \tag{4.18}$$

where L is the molecule's angular momentum and I is its moment of inertia about this axis,

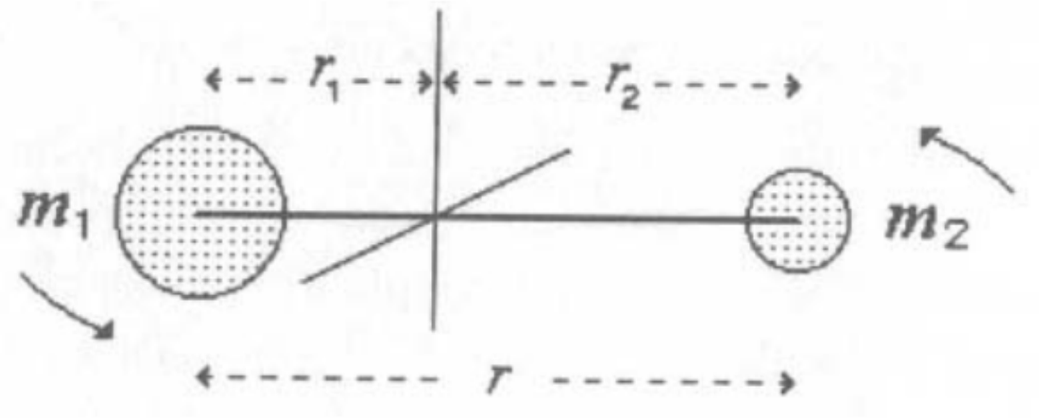


Fig 4.14 Schematic representation of a diatomic molecule such as HCI rotating around an axis through its center of mass.

The rotational energy eigenvalues of the molecule emerge from the quantization of its angular momentum.

The square, L^2, of the magnitude of the molecule's angular momentum obeys the quantization condition:

$$L^2 = J(J+1)\hbar^2 \tag{4.19}$$

where the rotational quantum number J can take the values J 0,1,2,3, ...

Substituting the allowed values of L^2 in equation (4.18) gives for the rotational energy eigenvalues (the rotational energy levels):

$$E_{rot} = \frac{J(J+1)\hbar^2}{2I} = J(J+1)Bh \tag{4.20}$$

where the constant B is defined by

$$B = \frac{h}{8\pi^2 I} \tag{4.21}$$

The difference, ΔE_{rot}, between two adjacent rotational energy levels, the J level and the J + 1 level, is:

$$\begin{aligned} \Delta E_{rot} &= \left[(J+1)(J+2) - J(J+1)\right]Bh \\ &= 2Bh(J+1) \end{aligned} \tag{4.22}$$

It follows, that the difference, ΔE_{rot}, increases by an amount 2Bh with each unit increase in the quantum number J, i.e., the differences between adjacent rotational energy levels increase linearly with the quantum number J.

The selection rule for radiative rotational transitions is $\Delta J = \pm 1$. Hence, the frequencies , f, of the photons associated with such transitions will be:

$$f = \frac{\Delta E_{rot}}{h} = 2B(J+1) \tag{4.23}$$

and the difference, Δf between the frequencies of successive spectral lines a constant:

$$\Delta f = 2B \tag{4.24}$$

This result is consistent with the second of the principle characteristics of the rotational spectra noted above, the equal spacing of the absorption lines along the frequency axis.

Bearing in mind that a system cannot absorb the photon corresponding to a 'jump' from a given state unless that state is already occupied, the abundance of lines in the rotational absorption spectrum indicates that a substantial number of rotational states, apart from the ground state, are extensively occupied under normal conditions.

Furthermore, the intensities of these lines indicate that these states are occupied to a comparable extent.

In a sample of gaseous molecules, the ratio between the number, N_J, of molecules occupying the J rotational state to the number, N_{J+1}, occupying the adjacent state J + 1 state is given by the Maxwell-Boltzmann distribution:

$$N_{J+1}/N_J = e^{-\Delta E_{rot}/kT} \tag{4.25}$$

where k is the Boltzmann constant and T is the sample's absolute temperature.

In the case of $HCl_{(g)}$, the constant B = 3.0x10^{11}Hz.

Substitution in equation(4.22) gives for the difference, ΔE_{rot}, between the energy of the J = 0 and J = 1 states:

$$\begin{aligned}\Delta E_{rot} &= 2Bh(J+1) \\ &= 2 \cdot 3.0 \cdot 10^{11} \cdot 4.14 \cdot 10^{-15} \cdot 1 \\ &= 2.5 \cdot 10^{-3} \text{eV}\end{aligned} \tag{4.26}$$

room temperature (25°C), kT = 2.6x10^{-2}eV.

Substituting these values in equation (4.25) gives:

$$N_{J+1}/N_J = e^{-2.5 \cdot 10^{-3}/2.6 \cdot 10^{-2}} \approx 0.9 \tag{4.27}$$

i.e., the J = 0 and J = 1 states are occupied almost to the same extent.

A similar calculation shows that even the J = 5 state has a population about 20% that of the ground state.

This explains the abundance of absorption lines in the rotational spectra, the first of the principle characteristics noted above.

We have examined only the rotational spectrum of diatomic molecules.

The spectra of polyatomic molecules are more complex but involve similar characteristics to those involved in the diatomic case.

4.2.3 Vibrational Spectra

Whereas diatomic molecules can vibrate in just one mode, to and fro along the axis joining their atomic centers, polyatomic molecules comprising n atoms have (3n - 6) vibrational modes.

If the molecule is also linear it has one more mode, i.e., (3n - 5) modes altogether (Fig 4.15).

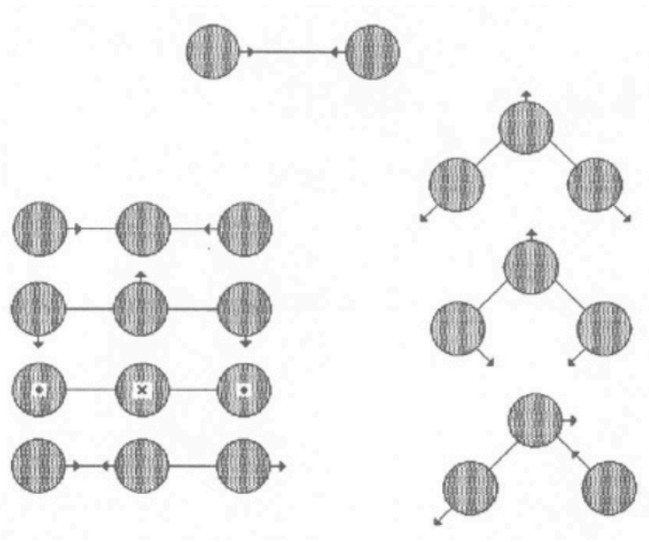


Fig 4.15 The vibrational modes of a diatomic and triatomic molecule.

At low energies, the vibrations of a diatomic molecule can be described in terms of a harmonic oscillator.

According to Planck's quantum hypothesis, the energy, E, of a harmonic oscillator of natural frequency f_0 can only take the values $E = nh f_0$ where $n = 0,1,2,3,.\ ...$

This implies that in its ground state, n = 0, the energy of the oscillator should equal zero and that the oscillator should be completely at rest.

However, this would contravene the Heisenberg Uncertainty Principle; complete rest implies no uncertainty whatsoever as to the system's position and linear momentum.

The correct expression for the energy, E_{vib}, of the oscillator is obtained from quantum mechanics.

The potential energy, U, of a simple harmonic oscillator is given by:

$$U = kx^2/2 \tag{4.28}$$

where k is the oscillator (spring) constant.

Substituting this expression for the potential energy in the one-dimensional time independent Schrodinger equation gives:

$$\frac{d^2\psi}{dx^2} + \frac{8\pi^2 m'}{h^2}\left(E_{vib} - \frac{kx^2}{2}\right)\psi = 0 \tag{4.29}$$

The energy eigenvalues given by the solution to this equation are

$$E_{vib} = \left(v + 1/2\right)hf_0 \tag{4.30}$$

where the oscillator's natural frequency, f_0 , is given by

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m'}} \tag{4.31}$$

The integer v, is called the vibrational quantum number and can take the values v = 0, 1,2,3, ……

Thus, in its ground state, v = 0, the oscillator has a 'zero-point' energy of $hf_0/2$ and so is never completely at rest.

At low energies, where the diatomic molecule can be regarded as a harmonic oscillator, the vibrational energy levels are equally spaced.

However, at higher energies, i.e., higher vibrational quantum numbers, the oscillator becomes anharmonic as is shown in Fig 4.16 and the energy levels close up.

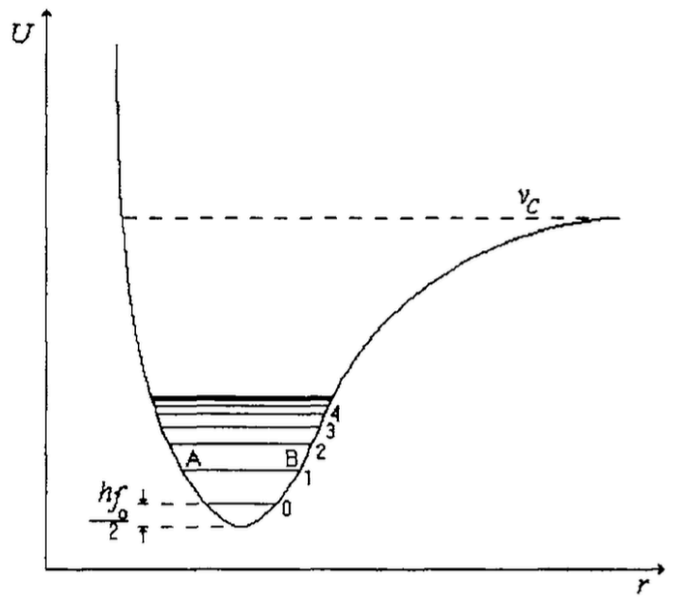


Fig 4.16 The potential energy U, of a diatomic molecule as a function of interatomic distance, r . The horizontal lines represent the vibrational energy levels, E_{vib}, of the molecule. At low energies the curve is parabolic, consistent with harmonic oscillations.

As the energy increases, so does the amplitude of the vibrations.

At the energy corresponding to the quantum number v, the bond linking the atoms breaks.

In practice, the vibrational spectra of diatomic molecules are very simple, usually comprising just a single strong line called the fundamental line.

There are two reasons for this:

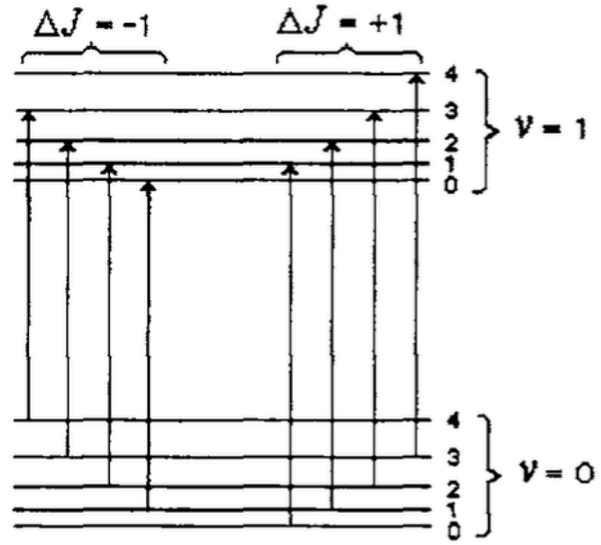
A. **The selection rule for radiative vibrational transitions is $\Delta v = \pm 1$. It follows, from equation (4.37), that only photons of frequency f_0 will be absorbed or emitted in transitions between low energy vibrational states.**

B. **A typical value for the difference between two adjacent vibrational energy levels is $\Delta E_{vib} = 0.5\text{eV}$. Consequently, at room temperature (25°C), the numerical value of the factor $e^{-\Delta E_{vib}/kT}$ in the Maxwell-Boltzmann distribution function is about 10^{-8}, i.e., the population of the excited states is negligible. All the molecules are in the ground state and so there are no anharmonic vibrations or transitions between higher states; only transitions between the $v = 0$ and $v = 1$ states take place.**

At low resolution, the fundamental line appears as a broad band.

This is because the transitions between the vibrational states are accompanied by transitions between rotational states (Fig 4.17).

Fig 4.17 Vibrational-rotational radiative transitions in a diatomic molecule. The transitions obey the selection rules $\Delta v = \pm 1$ and $\Delta J = \pm 1$.

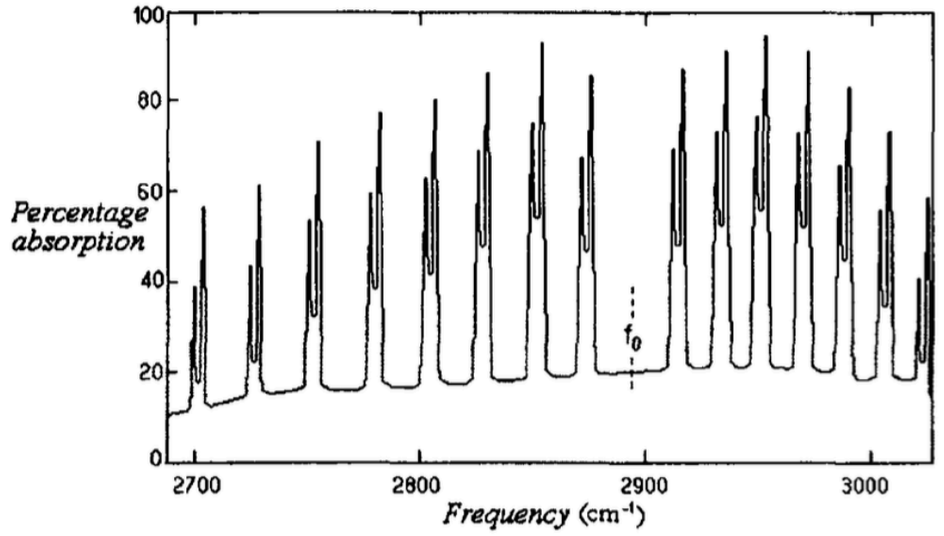


In these transitions, photons with energies in the range $E_{photon} = \Delta E_{vib} \pm \Delta E_{rot}$ where ΔE_{rot} is the energy involved in these rotational transitions, are absorbed or emitted by the molecule.

This range of photon energies shows up as a broadening of the fundamental line.

At higher resolutions, the band of the fundamental line is seen to comprise a series of fine lines, each corresponding to a rotational transition between different rotational sub-levels (Fig 4.18).

Fig 4.18 The &-red absorption spectrum of $HCl_{(g)}$ at high resolution. Each line that accords to a different rotational transition is itself split into two peaks, corresponding to the two isotopes C1-35 and C1-37 found naturally.



The frequencies, f, of these fine lines are given by:

$$f = f_0 \pm 2B(J+1) \tag{4.32}$$

Note that the $f = f_0$ line does not appear since it requires that $\Delta J = 0$ which contravenes the selection rule $\Delta J = \pm 1$.

A complete elucidation of the infra-red spectrum can only be obtained in the case of simple molecules.

Nevertheless, it is often possible to identify the presence of certain functional groups in polyatomic molecules by the approximate constancy of the frequency of the absorption band associated with them, irrespective of the compound in which the group appears.

For example, the -C = 0 carbonyl group vibrates with a frequency of 1706 cm^{-1} in acetone, 1715 cm^{-1} in acetaldehyde, 1663 cm^{-1} in acetic acid and 1736 cm^{-1} in methyl acetate (1 $cm^{-1} = 3x10^{10}Hz$).

This approximate constancy is the basis of the widespread use of infra-red spectroscopy in the identification of organic compounds and in the elucidation of their structure.

4.2.4 Electronic Spectra

Transitions between molecular electronic levels (ΔE_{elec} = 1- l0eV) are usually accompanied by transitions between the molecule's vibrational and rotational levels.

Accordingly, the frequencies of the photons emitted and absorbed in such transitions are given by:

$$f = \frac{\Delta E_{elec} + \Delta E_{vib} + \Delta E_{rot}}{h} \tag{4.33}$$

These frequencies are in the visible and the near ultra-violet.

Transitions between electronic states appear as a spectral band.

At higher resolutions these bands are seen to comprise closely packed lines corresponding to the various upper and lower vibrational and rotational levels.

All molecules, including homonuclear ones such as H_2 and N_2, exhibit electronic spectra since any change in the molecule's electronic configuration is always accompanied by a change in dipole moment.

An analysis of the fine structure of electronic spectra shows that the selection rule that applies to pure vibrational transitions, $\Delta v = \pm 1$, does not hold when the vibrational transition occurs together with an electronic transition.

This anomaly can be explained in terms of the following two general rules known as the F*ranck-Condon Principle*.

1. **The absorption of electronic energy occurs rapidly; a typical electronic 'jump' lasts just 10^{-16}s. On the other hand, bond vibrations are relatively sluggish, $\sim 10^{-13}$s. Thus, it can be assumed, the position and velocity of the atomic nuclei in the molecule do not change during an electronic transition.**
2. **The probability density in the ground state is greatest at the mid-point of the atomic nuclei's vibrations. Thus, it is from this point that transitions from the vibrational ground state, $v = 0$, will most probably occur. In other vibrational states, the probability density is greatest at the extremes of the vibration and it is from these points that the vertical lines representing the transitions from these states should be drawn—points A and B in Fig 4.16.**

Figure 4.19 shows the potential energy curves of the ground state, M, and an excited electronic state, M*, of a hypothetical molecule.

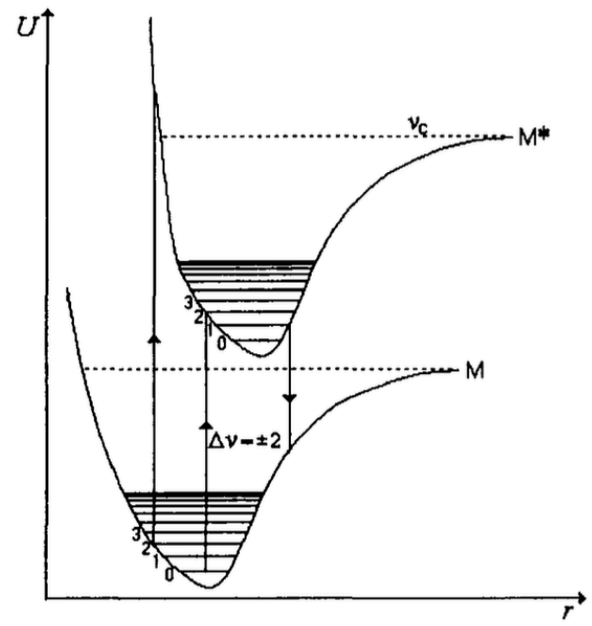


Fig 4.19 M and M* are the potential energy curves of the ground state and of an excited electronic state, respectively, as a function of the interatomic distance, r. It is apparent that a transition from one electronic state to another may well be accompanied by a transition between vibrational sub-states.

Transitions between electronic states can be represented by vertical lines on the potential energy diagrams.

Note that the dimensions of the excited state are different from those of the ground state; electronic excitation often affects molecular geometry.

Because the geometry of the two states changes only slowly, a molecule at the mid-point of the vibrational sub-state v = 0 in the electronic ground state can rise to the vibrational sub-state v = 2 in the excited electronic state, or can fall back from there into the ground state, even though both of these transitions are contrary to the selection rule $\Delta v = \pm 1$.

Similarly, the excitation of a molecule occupying an elevated vibrational sub-state in the electronic ground state can lead to the molecule's dissociation.

This occurs when the resulting vibrational sub-state in the excited electronic state has a quantum number greater than, v_c that which can be ascribed to the level asymptotic to the potential energy curve.

The electronic spectra of molecules composed of many atoms are very complex and are exceedingly difficult to analyze; their spectra are generally made up of a number of broad bands each corresponding to a particular electronic transition and each resolvable into numerous vibrational lines.

Almost all such molecules absorb ultra-violet radiation at wavelengths less than 150nm.

Colored materials, such as pigments and dyes, absorb and emit components of visible light, i.e., radiation of wavelengths in the range 400nm to 700nm.

This radiation corresponds to photon energies from 3.0eV down to 1.7eV.

These materials must therefore possess at least one electronic energy level that is between 1.7eV and 3.0eV above their electronic ground state.

This amount of energy is substantially less than the bond energy of a typical covalent bond.

Consequently, a material's coloration indicates that at least one of its electrons is relatively loosely bound and can be elevated to a higher level with comparative ease.

This electron is typically an unpaired electron or an electron in the π orbital associated with multiple bonds.

4.2.5 Raman Spectra

In 1928, Raman discovered that the light scattered by molecules contains frequencies other than those of the incident radiation; the phenomenon. is called Raman scattering.

The source of the additional frequencies is in the internal structure of the scattering particles.

Molecules both vibrate and rotate, and these motions can add to the vibrations induced by the electric field in the incident radiation.

The phenomenon is best understood in terms of photons.

As a first possibility, a photon of energy hf striking a molecule will be scattered without any change in its energy (frequency).

The molecule rises to a higher energy state from which it returns at once to its ground state by emitting an identical photon.

This is the mechanism of the normal Rayleigh scattering.

A second possibility is that the excited molecule emits a photon of lower energy (frequency) than that which struck it, using the surplus to enhance the energy of its vibrations and rotations.

In this case, the energy of the emitted photon is h(f - f_v) and the increase in the molecule's vibrational and or rotational energy is hf_v, (Fig 4.20).

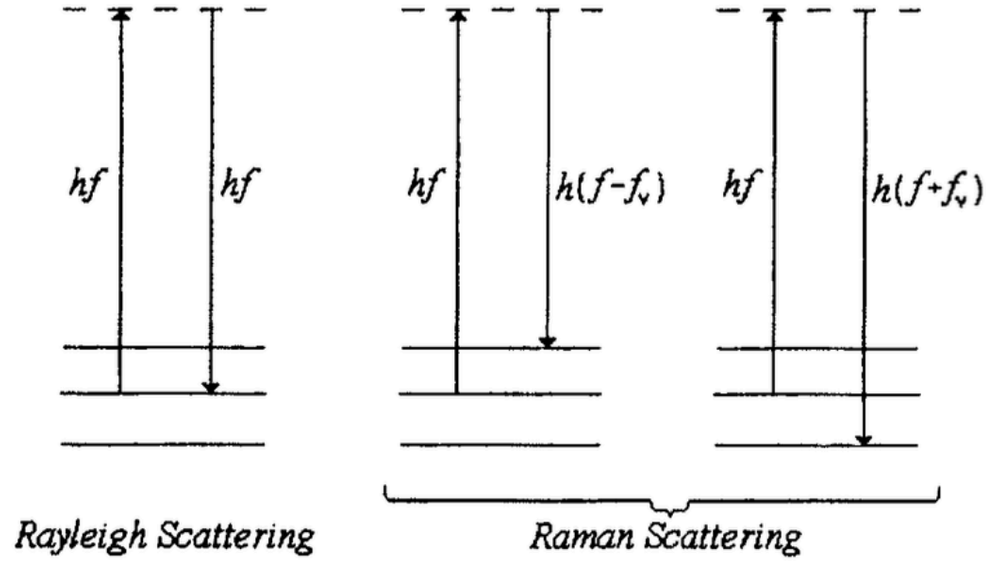


Fig 4.20 Raman and Rayleigh scattering interpreted in terms of the absorption and emission of photons. In Rayleigh scattering the photon emitted is identical to that absorbed. In Raman scattering the emitted photon can have a slightly higher or lower energy than that of the incident photon.

In rare instances, the incident photon will strike a molecule whose vibrational and or rotational energy has already been enhanced.

In this case, the molecule can emit a photon of energy h($f + f_v$). Thus, the scattered radiation can contain frequencies that are both slightly higher, $f + f_v$, and slightly lower, $f - f_v$, than that of the incident radiation.

Raman scattering is employed in the elucidation of molecular structures.

For example, the molecule N_2O has two possible configurations: NON and NNO .

The first configuration is symmetrical; the second is asymmetrical.

This difference in their structures expresses itself as a difference in their vibrational modes.

The Raman scattering of N_2O corresponds to vibrations of the asymmetrical configuration and thus the molecule's structure was determined.

Part 4.3

Fluorescence and Phosphorescence

When white light is passed through an aqueous solution of fluorescein, the transmitted beam is seen to have a red hue.

Evidently, the blue component of the incident light is absorbed by the fluorescein.

However, when viewed from the side, the same solution appears green.

Thus, though it absorbs blue light, it reemits green light (Fig 4.21).

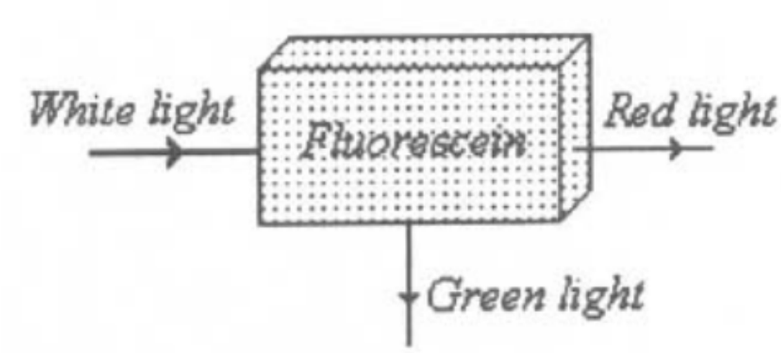


Fig 4.21 The passage of a beam of white light through a fluorescein solution.

The phenomenon is called fluorescence.

In general, the light emitted by a fluorescing material is of a longer wavelength than that it absorbs.

For example, in the case of fluorescein, $(\lambda_{green})_{Emitted} > (\lambda_{blue})_{Absorbed}$

When illuminated with ultra-violet radiation, certain naturally occurring minerals, such as fluorite and calcite, fluoresce and emit visible light.

Some washing powders contain fluorescent compounds that enhance the 'whiteness' of the laundered items in sunlight.

The familiar fluorescent lamp is a low pressure mercury vapor discharge tube whose inside walls are coated with a fluorescent material.

The bombardment of the mercury atoms in the tube by the accelerated electrons elevates them to an excited state.

As the atoms return to their ground state, they emit ultra-violet radiation, which impinges on the fluorescent coating on the inside of the tube.

This in turn emits white light.

Because fluorescent materials can be easily detected in the dark by the light they emit when irradiated with invisible ultra-violet radiation, fluorescence is widely used in research and industry, particularly for identification and labelling.

Fluorescence is readily explained in terms of molecular energy levels.

Whereas atomic electronic energy levels are fine and distinct, molecular energy levels are broadened and diffused as a result of the many vibrational and rotational sublevels they comprise.

Typical atomic and molecular energy level diagrams are compared in Figure 4.22,

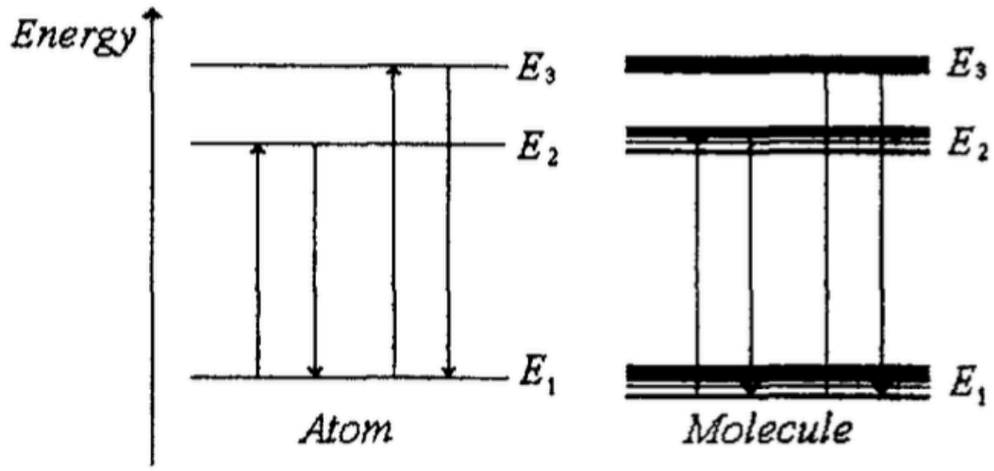


Fig 4.23 Hypothetical atomic and molecular energy level diagrams.A molecule in a fluorescent material that is excited to the E_2 or E_3 electronic state can lose energy in non-radiative transitions within the state (represented by the squiggly line) before emitting a photon and returning to the ground state, E_1

A molecule in an excited state can, just like an atom, return directly to its ground state by emitting a photon of the same frequency as that it absorbed.

However, a molecule has another alternative.

Before emitting the photon and returning to its ground state, it can lose some of its energy in collisions with other molecules in its environment, thereby dropping from a higher vibrational sub-level to a lower one within the same electronic level.

For example, consider a molecule in a liquid or a solid that absorbs a photon and rises from its ground state to the excited E_3 electronic state.

Whereas the average lifetime of the excited electronic E_3 state of the molecule is about 10^{-8} seconds, that of the vibrationa1 sub-states is just 10^{-12} seconds.

Because of the relatively long lifetime of the E_3 state, the probability that the molecule will return directly to the E_1 ground state is small.

During the time the molecule remains in this state it can lose vibrational energy as a result of collisions with the adjoining molecules.

In these collisions it falls from one vibrational state to another within the same electronic level.

It may even fall to the lower electronic level such as E_2 and drop from there to the ground state.

However. whatever happens, the energy of the photon the molecule emits will always be less than that of the photon it originally absorbed, i.e., the wavelength of the radiation the material fluoresces will be greater than that of the radiation it absorbs.

Fluorescence ceases when the radiation incident on the material is turned off.

However, some materials continue to glow even after the incident radiation has been extinguished; this phenomenon is called phosphorescence.

The essential difference between the phenomena is that whereas the total spin of the molecule is conserved in fluorescence transitions, in phosphorescence the excited states generally have a total spin different from that of the ground state.

In most instances, the ground state of a molecule is a singlet state, $S = 0$.

Absorbing a photon does not change the molecule's spin state.

However, in phosphorescent materials (generally solids) the non-radiative transitions that occur within the excited electronic state to which the system is elevated may result in its transition to a triplet state, $S = 1$.

For instance, whereas the E_3 level in Fig 4.24 is a singlet state the E_2 to which the system may descend as a result of collisions is a triplet state.

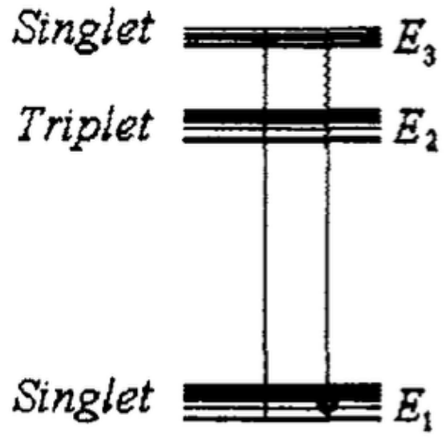


Fig 4.23 Phosphorescence. A system, excited from the singlet E_1 ground state to the E_3 state, falls by non-radiative transitions to the triplet E_2 state where it remains for a relatively long time because of the low probability of the 'forbidden' transition from a triplet to singlet state.

The probability of a transition from this triplet state back to the singlet state of the ground state is low, and so the lifetime of this intermediate triplet state is relatively long, from a thousandth of a second to even a few minutes.

Thus, the emission of the photon is delayed and the material fluoresces even after the incident radiation has been extinguished.

Part 5 Nuclear Physics

With a surface temperature of approximately 6000K, the sun must be emitting heat radiation at a rate of about $5x10^{23}$kW.

The mechanism by which the sun produces this immense amount of energy baffled scientists and astronomers for many years.

The great age attributed to the sun and its planets by cosmology, geology and the theory of evolution, precluded the possibility that it was either a chemical reaction such as combustion or that it resulted from gravitational collapse-the sun contracting under its own weight.

Calculations showed that neither of these processes could have sustained the required rate of energy production for the imputed aeons of deep time.

Becquerel's discovery in 1896 of the penetrating ionizing radiation emitted by substances containing the element uranium presented a similar problem.

Typically, the energies associated with this radiation are some million times greater than those associated with chemical reactions.

At first, Becquerel thought that the radiation was a form of phosphorescence and that its emission required the prior exposure of the uranium-bearing substances to sunlight.

However, he subsequently found that

> *... the radiations am emitted not only when the salts are exposed to light but also when they are kept in the dark, and for two months the same salts continued to emit these new radiations, without noticeable decrease in amount.*

Further investigations revealed that other heavy elements, such as thorium and radium, also emitted these radiations; they too are *radioactive*.

Becquerel remarked on the similarity between these radiations and the X-rays that had been recently discovered by Roentgen.

But the actual mechanism by which these radiations are generated and emitted from atoms remained an enigma for many years.

The nuclear model of the atom, proposed by Rutherford in 1913 to account for the scattering of α-rays by metal foils, had presented two fundamental problems: the stability of the atomic electrons and the incredibly great density it posited for the atom's nucleus.

The former, as we have seen, was subsequently explained by applying quantum mechanics to atomic structure.

The search for an understanding of the latter would not only resolve the enigma of the source of the sun's energy and of that emitted by radioactive substances, but would open up a whole new field of physics-the world of sub-atomic and high-energy particles-whose existence was unforeseen by either relativity or quantum mechanics.

Tragically, almost the first practical use made of this new knowledge was in the manufacture of weapons-nuclear weapons.

These horrific devices have presented mankind with the awesome means for its own almost total and instantaneous obliteration.

Now I am become Death, the destroyer of worlds - Robert Oppenheimer

However, wisely employed, this knowledge has the potential for developing virtually inexhaustible sources of useful energy: micro-suns working under human control, that could supply all of mankind's energy needs for as far into the future as anyone can possibly hope to plan.

Part 5.1 The Structure of the Nucleus

The proportions of atomic nuclei were first appraised by Rutherford in 1910 when he proposed his nuclear model of the atom.

His estimates were based on the results of the α particle scattering experiments he had initiated.

More recently, nuclear sizes have been measured by Hofstader.

The methods he used were based on the scattering of high energy electrons.

When the de Broglie wavelength of the electrons is of the order of the nuclear diameter, the scattered electrons produce a diffraction pattern from which the dimensions of the nucleus can be computed.

Nuclei are several orders of magnitude smaller than the atom.

Whereas atomic dimensions are of the order 10^{-10}m, those of nuclei are of the order 10^{-15}m to 10^{-14}m.

Nuclear dimensions are usually expressed in terms of the unit known as the *femtometer* or *fermi*:

$$\textbf{1femtometer (fm)} = 10^{-15}\textbf{m.}$$

In these units, $R_{Nucleus}$ = 1-l0fm where R represents the radius of a spherical particle.

Atomic nuclei are generally spherical in shape and of almost uniform density.

The positive charge of the atom is localized in its nucleus, together with almost all its mass.

Most nuclei possess both an intrinsic angular momentum, called nuclear spin, and a magnetic moment.

Typically, the magnitude of a nuclear magnetic moment is about a thousandth that of the electron.

However, there is no simple relationship between the spin of a nucleus and its magnetic moment; in some nuclei the magnetic moment is in the same direction as the spin and in others it is in the opposite direction.

Quantum mechanics must always be used in the description of nuclei and their interactions.

Accordingly, atomic nuclei occupy discrete quantum states and they absorb and emit energy by means of transitions between their energy levels.

5.1.1 Nucleons

Atomic nuclei are not elementary particles; they are themselves composed of more fundamental particles.

The particles that comprise atomic nuclei are termed nucleons; the term indicates a proton or a neutron.

Protons carry the unit positive charge, +e, and have a rest-mass some 1836 times greater than that of the electron.

Neutrons carry no net electrical charge; their rest-mass, 1839 times that of an electron, is marginally greater than that of the proton.

Protons and neutrons are both fermions; each has a spin of a half.

They also both possess a magnetic moment, the magnitude of which is about a thousandth that of the electron's magnetic moment.

However, whereas that of the proton is in the same direction as its spin, that of the neutron is in the opposite direction.

The existence of a magnetic moment in neutrons, despite the fact that they carry no net electrical charge, attests to an internal structure.

Neutrons are not electrically homogeneous.

Experiments on the scattering of neutrons have shown that they comprise an inner core of positive charge overlaid by an outer coating of negative charge.

Table 5.1: Properties of the electron and the nucleons

	Electron	Proton	Neutron
Rest-mass	$m_e = 9.11 \cdot 10^{-31}$ kg $= 0.00055$u	$m_p = 1836 m_e$ $= 1.00726$u	$m_n = 1839 m_e$ $= 1.00867$u
Radius	point	about 1fm	about 1fm
Charge	$-e = 1.6 \cdot 10^{-19}$C	$+e = 1.6 \cdot 10^{-19}$C	0
Spin	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
Magnetic moment	$-e\hbar/2m_e$	$+2.79e\hbar/2m_p$	$-1.91e\hbar/2m_p$

Like all minute particles confined to a particular region-a potential well-the nucleons in an atomic nucleus occupy discrete energy levels.

Two distinct, albeit similar, sets of nuclear energy levels exist, one for the protons the other for the neutrons each possess their own, albeit similar, set of energy levels.

Since the protons and neutrons are both fermions, Pauli's exclusion principle pertains to them.

Consequently, each nuclear energy level can hold at most a pair of protons or a pair of neutrons, with their spins opposed (Fig 5.1).

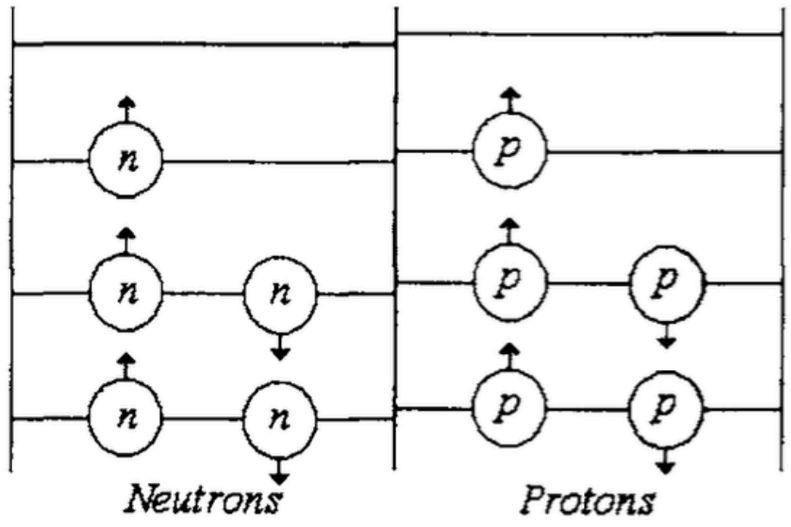


Fig 5.1 Nuclear energy levels. The proton energy levels are slightly higher than those of the neutrons because of their mutual electrostatic repulsion. Each energy level is occupied by no more than two particles with their spins opposed.

5.1.2 Nuclear Nomenclature

a. <u>Mass number</u>

The total number, A, of nucleons in an atomic nucleus, i.e., the number of protons plus the number of neutrons, is called its *mass number*.

The mass numbers of the naturally occurring nuclei range from 1 (hydrogen) to 238 (uranium).

It has been found, empirically, that, the number of nucleons in each unit volume of an atomic nucleus is ~0.15 nucleons/fm^3, irrespective of its mass number.

The following conclusions can be inferred from this observation:

(i) nucleons can be regarded as having a fixed volume;

(ii) the volume of a nucleus is directly proportional to its mass number;

(iii) the nuclear radius, R, can be estimated from the formula

$$R = R_0 \sqrt[3]{A} \tag{5.1}$$

where $R_0 = \sqrt[3]{3/4 \cdot \pi \cdot 0.15} \approx 1.2$ fm.

b. <u>Atomic number</u>

The number of protons, Z, in an nucleus is called its *atomic number*.

The atomic numbers of the naturally occurring nuclei range from 1 (hydrogen) to 92 (uranium).

The nuclear charge is equal to +Ze; in a neutral atom the number of electrons around the nucleus also equals Z.

The interaction of an atom with its surroundings-its chemical properties depends on its electronic structure.

This, in turn, depends *inter alia* on the number of protons in its nucleus.

Thus, all atoms having the same atomic number are atoms of the same chemical element.

For example, the atomic number of the element carbon is 6; ergo, all carbon nuclei contain six protons and possess a nuclear charge +6e.

c. <u>Neutron excess</u>

The number of neutrons, N, in an atomic nucleus is given by the formula

$$N = A - Z \tag{5.2}$$

The difference, N - Z , between the number of neutrons and the number of protons in a nucleus is called the *neutron excess*.

In stable nuclei, the general rule is that the larger the nucleus the greater the neutron excess (Fig 5.2).

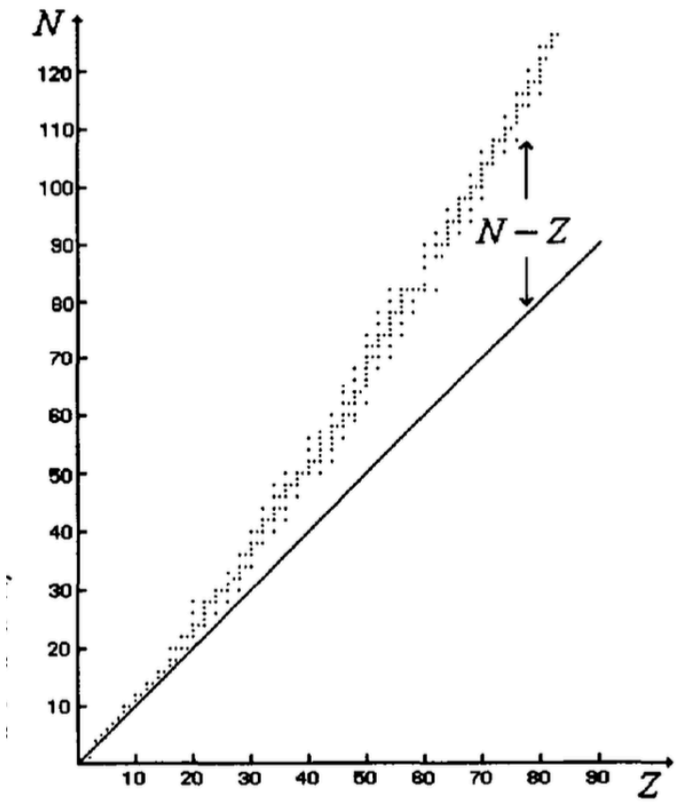


Fig 5.2 A plot of the number of neutrons, N, against the atomic number (number of protons), Z, in the 20 stable nuclei. The neutron excess, N - Z increases as the size of the nucleus grows.

d. <u>Nuclear composition</u>

The composition of a nucleus of the element whose chemical symbol is X, is designated by the symbol $\frac{A}{2}$X.

Thus, for example, the symbol $\frac{12}{6}$C designates a carbon nucleus containing 12 nucleons, 6 of which are protons, i.e., A = 12; Z = 6.

From equation (5.2), N = A - Z , it follows that this nucleus also contains 6 neutrons.

Alternative nomenclatures designate a nucleus by the symbols X-A and ^{A}X.

These give C-12 and ^{12}C, respectively, for the above carbon nucleus.

These nomenclatures omit the atomic number, Z, because it is implicit in the element's symbol.

5.1.3 Nuclear Masses; Isotopes

Nuclear masses can be measured with great accuracy with the help of techniques such as mass spectography.

They are variously expressed in terms of different units:

* **the kilogram, kg;**
* **the atomic mass unit, u:**
* **the relativistic unit, MeV/c^2 .**

Table 5.3: The rest masses of the electron and the nucleons in various units

	kg	u	MeV/c^2
Proton	$1.6726 \cdot 10^{-27}$	1.00726u	938.3
Neutron	$1.6750 \cdot 10^{-27}$	1.00867u	939.6
Electron	$9.109 \cdot 10^{-31}$	0.00055u	0.511

The atomic mass unit is defined as one twelfth of the mass of the carbon atom that contains 6 protons, 6 neutrons and 6 electrons, i.e., the isotope C-12.

Given that, by definition, the mass of a mole of such carbon atoms is 0.012kg, this gives:

$$1\text{u} = \frac{0.012}{12 \cdot 6.022 \cdot 10^{23}} = 1.661 \cdot 10^{-27}\,\text{kg} \tag{5.3}$$

From the mass-energy relationship $E = mc^2$, the energy equivalent, E_{1u}, of one atomic mass unit is:

$$E_{1u} = m_{1u}c^2 = 1.66 \cdot 10^{-27} \cdot 9 \cdot 10^{16} = 1.49 \cdot 10^{-10}\,\text{J} \tag{5.4}$$

Or, alternatively, in units of millions of electron-volts:

$$E_{1u} = \frac{1.49 \cdot 10^{-10}}{1.6 \cdot 10^{-19} \cdot 10^{6}} = 931\text{MeV} \tag{5.5}$$

Though all the nuclei of the atoms of a particular element contain the same number of protons, they do not necessarily contain the same number of neutrons, i.e., nuclei of the same element can have different mass numbers.

Atoms of the same element that possess different mass numbers are called *isotopes*.

For example, the element uranium has a number of isotopes, the principle ones being U-235 and U-238. Both have 92 protons in their nuclei.

But, whereas the U-235 isotope has 143 neutrons in its nucleus, that of the U-238 isotope contains 146 neutrons.

The isotope of hydrogen with mass number A = 2 is called *deuterium* or *heavy hydrogen*.

The nucleus of this isotope is called a *deuteron* and is designated ^{2}D or ^{2}H; it is composed of one proton and one neutron. Heavy water, which constitutes about 0.015% of naturally occurring water, embodies this isotope of hydrogen.

Table 5.2: The naturally occurring isotopes of Hydrogen and Chlorine

	Atomic Number	Symbol	Number of protons (Z)	Number of neutrons (N)	Mass Number (A)	Atomic Mass	Relative Abundance (%)
Hydrogen	1	^{1}H	1	0	1	1.008u	99.99
		^{2}H	1	1	2	2.014u	0.01
		^{3}H	1	2	3	3.016u	<0.001
Chlorine	17	^{35}Cl	17	18	35	34.97u	75.53
		^{37}Cl	17	20	37	36.97u	24.47

Because all the isotopes of a given element exhibit the same chemical properties, they cannot be separated by chemical means.

The traditional methods of separating isotopes exploit the differences in their masses.

These methods have included mass spectrography, centrifuging and processes that depend on the different rates at which heavy and light particles diffuse.

More recent techniques using tunable lasers rely on the different nuclear excitation energies of the isotopes.

5.1.4 Nuclear Binding Energy

An examination of nuclear masses shows that the mass of a nucleus is always less than that of its constituent nucleons.

For example, the mass of a $^{12}_{6}C$ nucleus, $m_{^{12}_{6}C} = 1\,1.997u$, is less than that of 6 free protons and 6 free neutrons:

$$6m_p + 6m_n = (6 \cdot 1.00726u) + (6 \cdot 1.00867u) = 12.0956u$$

The difference,

$$\Delta m = 6m_p + 6m_n - m_{^{12}_{6}C} = 0.0986u$$

between the mass of the nucleus and that of its constituent nucleons is called the *mass defect* of the nucleus.

$$E_B = 931 \Delta m \text{ MeV} \tag{5.6}$$

Alternatively, this difference can be expressed as an *energy defect*, E_B, such that, when Δm is expressed in atomic mass units,

This quantity is also termed the binding energy of the nucleus.

It is the amount of energy that would be emitted during the formation of a nucleus from its constituent nucleons.

Alternatively, it can be defined as the quantity of energy that must be invested in order to break up a nucleus into its free constituent nucleons.

A key parameter in the study of atomic nuclei is their *average binding energy per nucleon*, E_B/A.

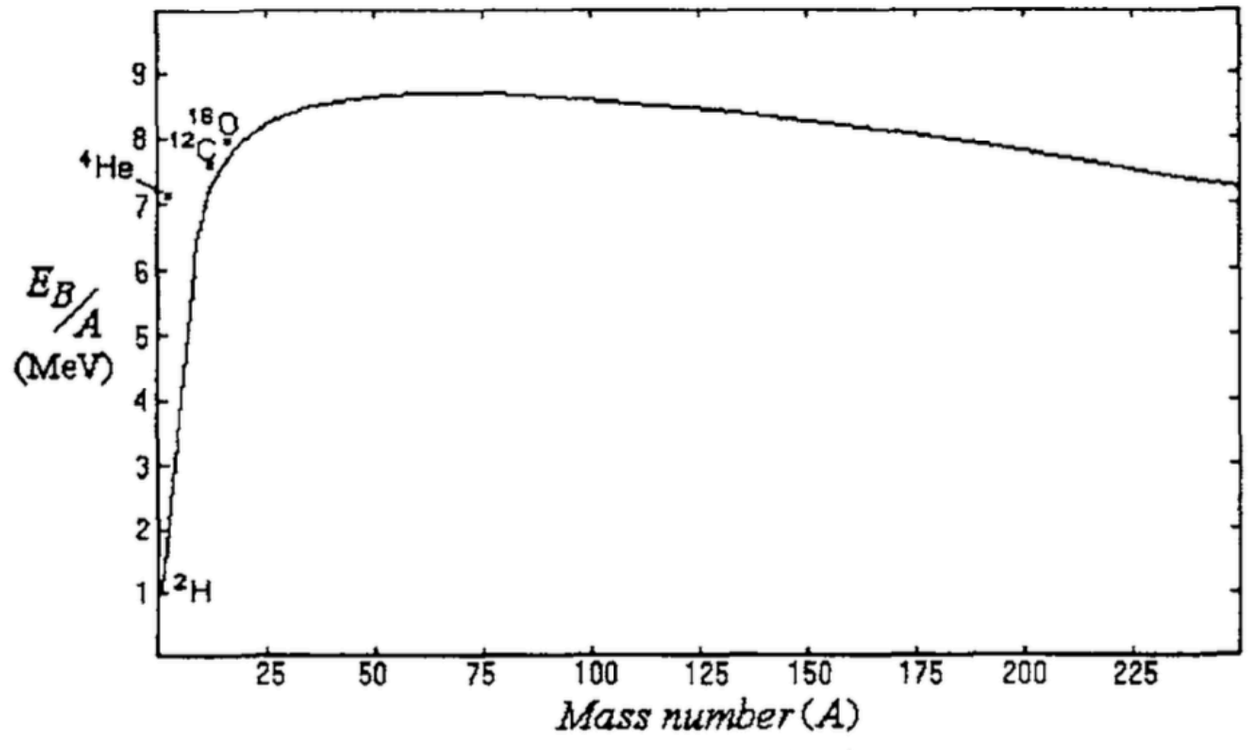
In the case of the $^{12}_{6}C$ nucleus, its value is

$$E_B/A = 91.8/12 = 7.65\,\text{MeV} \tag{5.7}$$

The average binding energy per nucleon varies from one nucleus to another.

It reaches its maximum value, almost 9MeV, in the nuclei of the elements situated just before the middle of the periodic table, i.e., for values of A between 40 and 100 (Fig 5.3).

Fig 5.3 The average binding energy per nucleon, E_B/A, as a function of the mass number, A. It reaches its maximum in the range A = 40 to 100.



5.1.3 The Nuclear ('Strong') Force

Atomic nuclei are very small; their diameters range from about l fm to 8 fm.

Furthermore, all atomic nuclei, except for those of hydrogen, contain two or more protons.

At such close distances, the electrostatic repulsive force that acts between two protons is of macroscopic proportions-tens or even hundreds of newtons.

This force, acting between the closely packed protons, should cause the nuclei to blow themselves apart.

The fact that atomic nuclei do not spontaneously explode, despite the presence of these potentially destructive forces within them, requires an explanation.

In order to account for the existence of atomic nuclei despite the mutual repulsion of their constituent protons, it is postulated that a short-range attractive force acts between nucleons: a cohesive force that is substantially stronger than the repulsive electrostatic force between the protons.

This second force, called the nuclear force or the strong force, binds the nucleons in a nucleus together.

At very close distances, this tremendous force of attraction acts between all the nucleons: protons attract protons, neutrons attract neutrons and protons attract neutrons.

The strong force is essentially charge independent: the forces that acts between two protons, between two neutrons or between a proton and a neutron have virtually the same strength.

Evidence for this is provided by the binding energies of 'mirror nuclei'.

Pairs of nuclei, such as ^{3}He and ^{3}H, ^{7}Be and ^{7}Li or ^{35}Cl and ^{35}Ar, that have the same mass number, A, and can be obtained from one another by the conversion of a neutron into a proton, or vice versa, are called mirror nuclei.

Such pairs of nuclei are found to have almost identical binding energies.

Evidently, the electrical charge carried by the nucleons has no significant effect on their binding energies.

The slight differences that are found are attributed to the energy of the electrostatic repulsion that acts between the protons in the nucleus.

The nuclear force is very short-range.

Plainly, its reach does not extend to distances of the order of magnitude of the atom, ~10^{-10}m.

If it did, the isotopes of a given element would not all exhibit the same chemical properties.

In practice, at distances beyond 3 fm it fades away rapidly.

Thus, although an attractive force of some 2000 newtons acts between two nucleons at a distance of 2 fm, nucleons on one side of a large nucleus such as the uranium nucleus, are not directly affected by the nuclear force exerted by the nucleons on the opposite side.

Experimental evidence of the existence and nature of the nuclear force has been obtained from nuclear scattering experiments of the kind carried out by Rutherford.

When such experiments are carried out with very high energy protons or α particles, qualitatively different results from those he encountered are obtained.

At particle energies of just a few MeV, such as those Rutherford employed, the scattered protons or α particles are unaffected by the nuclear force; even their closest encounters with the nucleus do not bring them within its range.

However, at higher energies the particles can overcome the electrostatic forces of repulsion acting between them and the nucleus, and can penetrate so deep into the atom that they 'touch' the nucleus and enter the range of the nuclear force.

The increase in the neutron excess, N - Z, in stable nuclei as their size increases (Fig 5.2) and the variation of the average binding energy per nucleon, with mass number (Fig 5.3) can both be qualitatively explained in terms of the respective ranges and strengths of the nuclear and electrostatic forces, as follows.

* **In the smallest nuclei, those with a mass number of ten or less, $A \leq 10$, all the nucleons are within range of the nuclear force exerted by all of the other nucleons. The addition of a nucleon to such a nucleus enhances the overall cohesive force binding the nucleons. As a result, the average binding energy per nucleon rises sharply with each increase in mass number (Fig 5.3). In these small nuclei, the ratio between the number of neutrons to the number of protons is close to one (Fig 5.2), and the effect of the repulsion between the protons is negligible.**

- **As the mass number, A, rises above 10, each additional nucleon produces a diminishing return in the amount by which the average binding energy per nucleon increases. This is because the additional nucleons are only attracted by those closest to them, not by all the nucleons in the nucleus. Moreover, as the mass number increases and the nucleus grows larger, the number of nucleons 'in touch' with each additional nucleon drops lower still. The average average binding energy per nucleon reaches its maximum value at around $A \approx 60$.**

* Above $A \approx 60$, the average binding energy per nucleon starts to decline. This is attributed to the repulsive forces between the nuclear protons. Although the number of nucleons 'in touch' with each additional nucleon ceases to rise, all the extra protons are in range of the electrostatic forces of repulsion exerted by all the others. As a result, whereas the average cohesive force remains unchanged, the repulsive forces within the nuclei continue to grow. This is moderated to some extent by the increase in the neutron excess, $N - Z$, as the nuclei become larger; the extra neutrons add cohesion without contributing any repulsion and by stretching the average distance between the protons they weaken the repulsive forces between them. Thus, stable heavy nuclei can exist despite the large number of protons they contain. The heaviest stable nucleus is ^{209}Bi; the ratio between the number of its neutrons and protons being $N/Z = 1.52$. Heavier naturally occurring nuclei do exist but they are all unstable, i.e., radioactive.

In 1935, Yukawa suggested that the interaction between nucleons could be described in terms of a potential energy function, U_N, such that the total potential energy of the interaction between two protons is given by the sum, $U = U_N + U_E$, of the nuclear and electrostatic interactions (Fig 5.4).

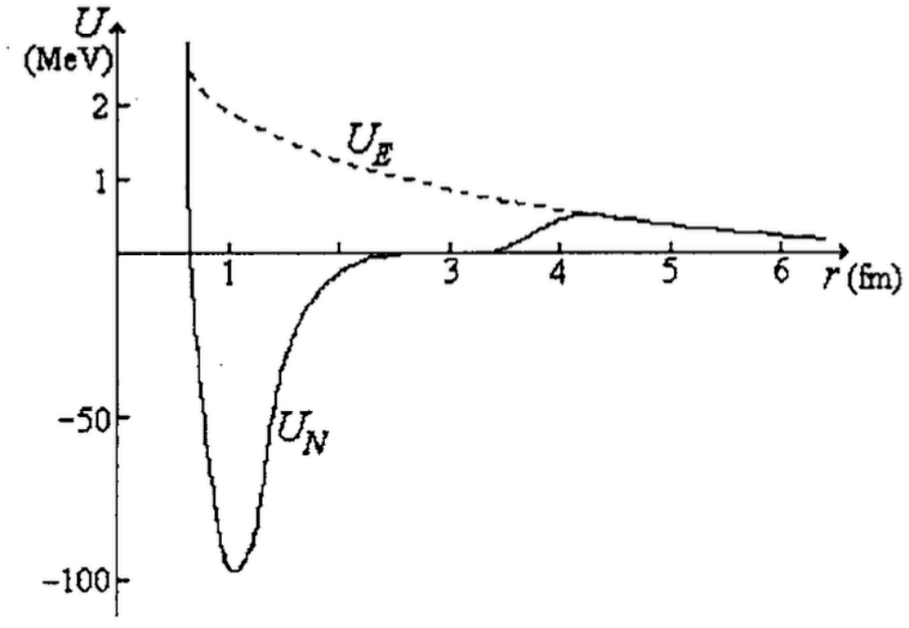


Fig 5.4 The total potential energy, U, of the interaction between two protons as a function of the distance, r, between them. U_E is the electrostatic potential energy and UN is the nuclear potential energy; $U = U_E + U_N$. Note that the energy scale is different for positive and negative energy values. Where $r > 4$ fm the total potential energy curve, U, merges with that of the electrostatic potential energy, U_E, and where $r < 3$ fm it merges with that of the nuclear potential energy, U_N.

Since nucleons can be regarded as particles of fixed volume, when they are closer than 0.8 fm to one another, a repulsive force, greater than the attractive nuclear force, acts between them; this appears as positive energy in the potential energy curve.

At separations greater than 4 fm, the nuclear force is out of range; only the electrostatic repulsion between the protons operates and so the total potential energy is positive.

At separations between 0.8 fm to 3 fm, the cohesive nuclear force between the protons overwhelms the electrostatic repulsion between them; the net force between them is attractive and so the total potential energy is negative.

However, in reality, the nuclear force is too complex to be summarized in terms of a simple function of the distance between the nucleons.

Experiments show that the nuclear force also depends on the spin of the nucleons and does not necessarily act along the axis joining the particles.

Despite all the knowledge we have about nuclear forces, the function that correctly describes nuclear interaction energy has yet to be discovered.

5.1.6 Nuclear Models

In the attempt to understand the structure and stability of nuclei, three models have been proposed:

the Liquid Drop model, the Shell model and the Collective model.

The Liquid Drop model, first proposed by Niels Bohr, views the nucleus as a droplet of nuclear fluid.

It assumes that each nucleon interacts only with its closest neighbors in the nucleus, like the molecules in a liquid.

By analogy with liquid drops, the model predicts the generally spherical shape of nuclei and provides a good qualitative description of the fission of large nuclei.

Quantitatively it provides a function for the binding energy per nucleon in terms of the mass number that closely fits the empirical data.

The Shell model attempts to apply the same quantum tools to nuclear structure as were successfully applied to the elucidation of the electronic structure of atoms.

It regards each nucleon as interacting with the overall force field created by the other nucleons comprising the nucleus.

Its outstanding achievement is a computation of the nucleonic energy levels that gives a rationale to the postulate that the nucleons are configured in shells inside nuclei.

In making the calculation it was assumed that the nucleons possess an orbital angular momentum, **L**, in addition to their spin, **S**, the two being coupled to form a total nuclear angular momentum, **I**, whose quantization prescribes the nuclear energy levels and hence the nucleonic configuration.

The Collective model is eclectic, uniting the best features of the other two.

Not surprisingly, it has been the most successful of the three.

Nevertheless, it still fails to give a comprehensive picture of the nucleus.

The Collective model views the nucleus as comprising a stable core of closed shells, each containing a magic number of nucleons, with which the 'extra' nucleons in the nucleus, i.e., those that are not constituents of the core, interact.

These 'extra' nucleons occupy quantum states whose energies are determined by their interaction with the core.

The core in tum is agitated by this interaction.

As a result, rotational and vibrational motions are set up in the nucleus that are analogous to those that occur in droplets.

5.1.7 The Elementary Particles of Matter

The notion that the atomic nucleus comprises yet smaller particles, begs the question of whether these particles -the nucleons- might not themselves be composites of still smaller particles.

And if so, where does the sequence (the 'pecking order') end?

We shall conclude our discussion of the structure of the nucleus by briefly considering some of the elementary building blocks from which matter is believed to be constructed.

a. **Pions**

As well as proposing a mathematical function to describe the nuclear force, Yukawa also suggested a mechanism for its action.

He posited, that the strong nuclear force that acts between nucleons arises from an exchange of particles between them.

Originally called π *mesons* these particles are now called *pions*.

The mechanism by which they mediate the strong interaction between nucleons is illustrated in the Feynman diagrams in Fig 5.5.

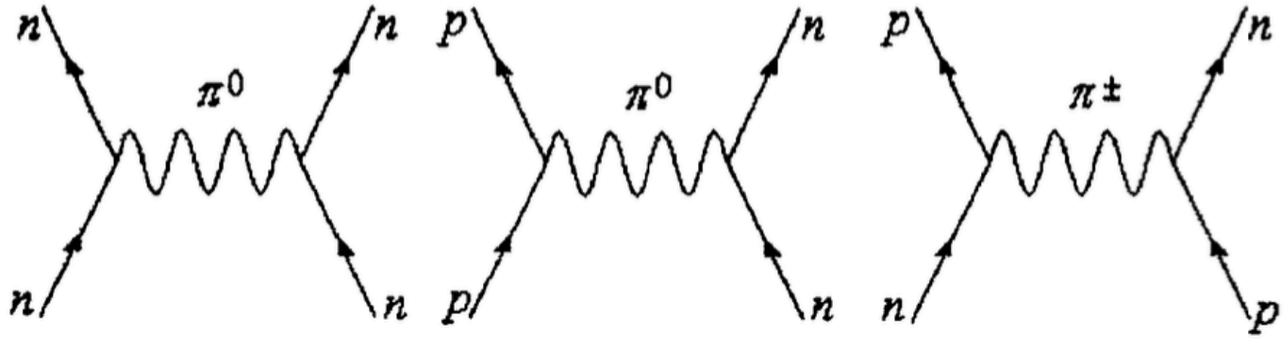


Fig 5.5 Feynman diagrams showing the strong interaction between nucleons mediated by virtual pions. The neutral pion, π^0, mediates interactions in which the nucleons retain their identity. The charged pions, $\pi^{\pm}$, mediate those in which protons become neutrons or neutrons become protons.

The notion that particle-particle interactions are mediated by an exchange of other particles is not new to us.

It is a basic concept of the quantum theory of fields.

The mediating particles can be actual particles as when two atoms form a covalent bond by the exchange of electrons or they can be virtual particles as in the interaction between stationary charged particles.

The spontaneous generation of a virtual particle, in the absence of any external source of energy, is allowed for by Heisenberg's uncertainty principle: the particle can appear on condition that it exists for a period no longer than $\Delta t \approx \hbar/\Delta E$, where ΔE is its total energy.

If it is to mediate the interaction between two particles, the ephemeral particle must manage to move from one to another during its fleeting lifetime.

Evidently, the less time it has to achieve this, the more energy must be 'borrowed', i.e., the heavier must be the particle.

Given that the equilibrium distance between two nucleons is typically $r \approx 1.7\text{fm}$ and assuming that the virtual particles that mediate the strong interaction between them move at close to the speed of light, c, their lifetime can be estimated to be

$$\Delta t = \frac{r}{c} = \frac{1.7 \cdot 10^{-15}}{3 \cdot 10^{8}} = 0.57 \cdot 10^{-23}\,\text{s} \tag{5.8}$$

This gives for the particle's total energy:

$$\Delta E = \frac{\hbar}{\Delta t} = \frac{6.58 \cdot 10^{-22}}{0.57 \cdot 10^{-23}} = 115.4\,\text{MeV} \tag{5.9}$$

This is equivalent to a rest-mass of 0.124u, i.e., somewhere between the mass of an electron and that of a nucleon.

Virtual pions cannot be directly observed.

In order for them to become observable, an external supply of energy must be provided.

According to the above calculation, the realization of free pions requires the investment of considerable energy; at least 115.4MeV.

To furnish a stationary target particle with this much energy in a collision, the incident particle must possess considerably more.

Before the construction of powerful particle accelerators in the 1950's the only source of such particle energies was cosmic rays.

Free pions were first observed by Powell in 1947 in the microscopic tracks they left in nuclear emulsions exposed at high altitudes to the bombardment of cosmic rays.

Three different pions were discovered: a neutral pion, π^0, and two charged pions, π^+ and π^-.

All three pions are unstable: the neutral pion decays after a lifetime of ~10^{-17}s into a pair of high-energy photons-y rays; the charged pions, π^+ and π^-, decay after a lifetime of ~10^{-8}s, by stages, into a positron, e^+, and an electron, e^-, respectively.

b. <u>Particles and anti-particles</u>

The existence of the positron or *anti-electron* was predicted by Dirac in 1928.

It arose as a possible solution of the wave function of the electron when relativistic considerations were incorporated into its formulation.

The solution to the wave equation required the existence of negative energy states.

But if such states existed and were unoccupied, free electrons possessing positive energy would be expected to fall into them, emitting a photon in the process.

Dirac postulated that these states are all generally occupied, even in a vacuum.

However, should one be unoccupied for any reason, a hole would be left in the 'sea' of filled negative energy states.

This hole, which behaves like a particle with positive charge, appears to us as a positron.

The positron, e^+, was first identified by Anderson in 1932 in cosmic rays.

Positrons and electrons are identical except for their electric charge; the positron is the anti-particle of the electron.

The charged pions are also identical except for the sign of the electric charge they carry; the negatively charged pion is the anti-particle of the positively charged one.

It transpires, that for every particle there exists a corresponding anti-particle.

In general, the particle and its anti-particle have the same mass, spin and lifetime, if they are unstable, but opposite electromagnetic properties.

If the one is positively charged the other will be negatively charged; if the one has its magnetic moment in the direction of its spin, that of the other will be in the opposite direction.

Chargeless particles like the neutral pion and the photon are their own anti-particle.

The anti-proton, p^- or $\bar{p}$, was discovered in 1955 in the shower of particles produced when a target was bombarded with protons that had been accelerated to an energy of 6.2GeV (6.2×10^3 MeV).

There is no obvious reason why matter composed entirely of anti-particles should not exist.

Such 'anti-matter' would behave exactly like ordinary matter.

However, if matter and anti-matter were to meet, they would mutually annihilate with the emission of a flood of very high energy photons, γ rays.

No such emissions have been observed on earth or anywhere in space.

For whatever reason, there does not appear to be much anti-matter around.

c. <u>Particle physics</u>

The discovery of the free pions in cosmic ray tracks in 1946 marks the birth of modern particle physics.

Subsequent investigations have revealed that pions and anti-protons are not the only species produced when high energy particles are slammed into stationary targets or into one another.

A plethora of unstable and short-lived is particles (more than 300) has been observed in the debris produced when protons or electrons accelerated up to energies of 1000GeV (106MeV) are smashed into stationary targets or when two such high-energy particles, or a particle and its anti-particle, collide head on.

The early particle accelerators were linear.

A supply of charged particles was accelerated along the length of an evacuated tube by the electric field produced by large static electric charges.

Most modem particle accelerators are descended from the cyclotron invented by Lawrence in 1931.

This ingenious device employs magnetic fields to keep the charged particles moving in circular paths whilst an alternating electric field of moderate strength accelerates them incrementally (Fig 5.6).

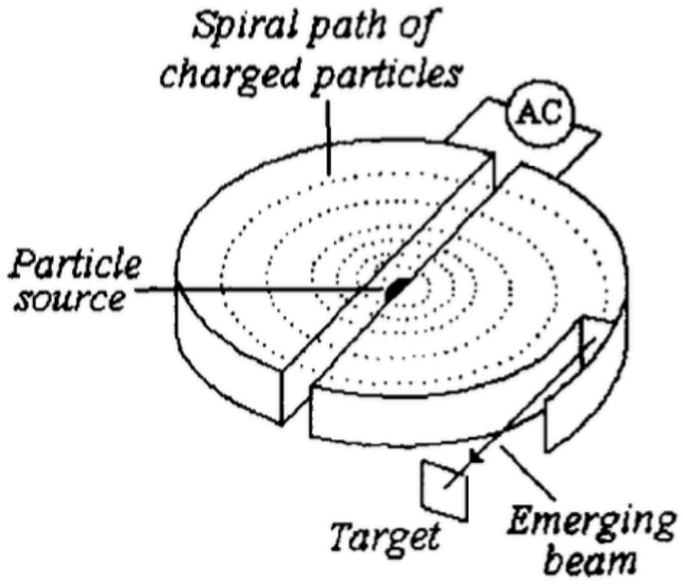


Fig 5.6 The cyclotron. A strong vertical magnetic field holds the charged particles (ions) on a circular path within two flat horizontal semicircular metal vessels---'Dees'. An alternating potential of the same frequency as that of the orbit is applied across the gap between the Dees so that the charged particle is given a push every time it crosses the gap, accelerating it incrementally to higher and higher speeds. As the speed increases, the particles spiral outwards until they emerge and strike the target.

At the very high particle velocities achieved in these devices, allowance must be made for relativistic effects such as the greatly increased mass of the particles as they approach the speed of light.

As a result, in order to reach the tremendous energies the researches wish to investigate, gigantic and expensive machines are required.

A proton accelerator at the Fermilab near Chicago has a radius of l km and produces beams with an energy of l000GeV (a speed of 0.9999995c), and the Super Proton Synchrotron at the Centre European de la Recherche Nucleaire (CERN) near Geneva has a radius of 1.2 km.

The biggest accelerator at CERN is the Large Hadron Collider-LHC which measures 27 km around and is housed in a tunnel about 100m underground- 13.0 TeV.

The particles created in the high-energy collisions are detected either in counters such as scintillation counters that register the passage of charged particles or in track-recording devices such as bubble chambers that provide pictures of the paths followed by the particles.

d. <u>The Quark Hypothesis</u> (before Standard Model development)

Attempts to make sense of the bewildering host of unstable particles produced from these brute-force high energy collisions and of the processes by which the particles interact and decay, have centered on the notion that particles such as the protons and neutrons that constitute atomic nuclei, as well as the pions that mediate between them, are themselves composed of still more fundamental particles.

The evidence for the existence of an internal structure in these particles is well established.

In experiments similar in their design to that of the Franck-Hertz experiment, evidence has been found for the existence of excited nucleon states, i.e., well-defined higher states to which a nucleon can been raised by the absorption of specific amounts of energy.

In these experiments, a beam of accelerated protons is aimed at a liquid-hydrogen target and the distribution of the energies of the scattered protons is analyzed.

The distribution indicates that although most of the incident protons collide elastically with those in the target, some do so inelastically.

In these inelastic proton-proton collisions, specific amounts of energy are taken up by the target protons.

Similar results are obtained when protons are bombarded by energetic π^+ pions.

The result of the interaction between the π^+ pion and the proton is the creation of three new pions.

$$\pi^{+*} + p \rightarrow \pi^+ + p + \left(\pi^+ + \pi^- + \pi^0\right)$$

In this case too, an analysis of the distribution of the energies of the three new pions indicates that the reaction can proceed via an intermediate stage, usually designated η_0, that is characterized by a specific energy, ~550MeV.

This intermediate stage, which is usually termed a *resonance*, can be regarded as being a very-short-lived particle, ~10^{-20}s, such that overall the reactions can be described as occurring in two steps:

$$\pi^{+*} + p \rightarrow \pi^+ + p + \eta_0 \qquad\qquad \eta_0 \rightarrow \pi^+ + \pi^- + \pi^0$$

At different times, various models have been suggested for the inner structure of the nucleons.

For many years, they were considered to comprise an inner core surrounded by a cloud of pions; a sort of microcosm of the nuclear model of the atom.

More recently, it has been suggested that they are composed of two different fundamental components called the up quark and the down quark.

According to this hypothesis, protons contain two up quarks and one down quark; neutrons, two down quarks and one up quark (Fig 5.7).

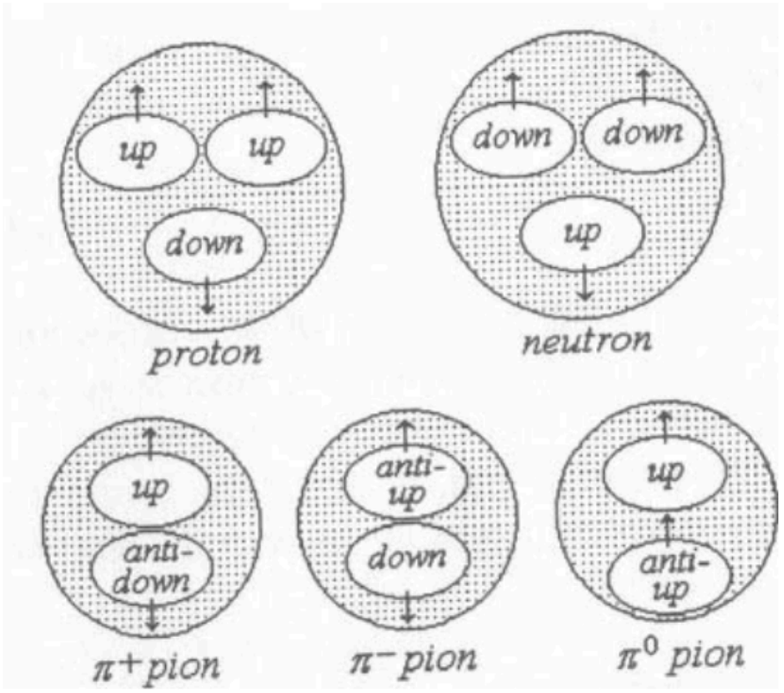


Fig 5.7 The quark configurations of the nucleons and pions. The arrows indicate the spins. The composite particle's spin and charge equals that of all its components.

It is postulated that the quarks carry fractional electric charges.

The up quark carries a charge of $+\frac{2}{3}e$; the down quark a charge of $-\frac{1}{3}e$.

Corresponding to each quark. there is a anti-quark. that carries an electric charge of the opposite sign.

Furthermore, it is suggested that it is the interactions between the quarks that underlie the strong force.

Quarks are presumed to be truly elementary particles, i.e., essentially point particles that, like electrons, possess no internal structure.

They are also fermions. However, there the similarity ends, for, unlike the quarks, electrons do not participate in strong interactions.

Table 5.4: The component elementary particles of ordinary matter

Particle	Type	Rest Mass[20] (MeV/c^2)	Spin	Charge (e)
Up	quark	~300	½	$+\frac{2}{3}$
Down	quark	~300	½	$-\frac{1}{3}$
Electron	lepton	0.511	½	−1

Whether they participate in the strong interaction, or do not, is the basis of the division of elementary particles into two distinct groups: leptons and hadrons:

> **Particles that do not take part in strong interactions are called leptons: electrons are leptons.**
>
> **Particles that do participate in strong interactions are called hadrons: protons and neutrons are hadrons.**

The hadrons are further sub-divided into baryons and mesons, according to the number of quarks they contain (Fig 5.7).

Protons and neutrons each contain three quarks; they belong to the family of particles called the baryons, each member of which is a fermion.

Pions are composed of a quark and an anti-quark; they belong to the family of particles called the mesons, each member of which is a boson.

Together, the baryon and meson families make up the clan of the hadrons.

Unlike the protons, neutrons and electrons that comprise atoms and nuclei, free quarks have never been observed; all quarks appear to be locked inside the particles they comprise.

Similarly, the photon-like particles called gluons that, it is postulated, mediate the interactions between the quarks have never been found experimentally.

Nevertheless, the quark hypothesis has proved a useful and effective model for explaining both nuclear processes and the products of collisions between high-energy particles.

Evidence for the presence of hard point-like constituents in nucleons has been obtained from the scattering of very high energy electrons (short De Broglie wavelength, ~I0^{-17}m) by protons and neutrons.

The results indicate that the nucleons' constituents have fractional charges and spin 1/2.

However, they also indicate that only part of the nucleon's momentum is stored in its quarks; part is also stored in electrically neutral particles within the nucleon, possibly the gluons.

Remarkably, the bewildering host of unstable particles produced from the brute-force high energy collisions that provide the experimental data of particle physics and from, the processes by which the particles interact and decay, can be systematized.

This requires postulating the existence of additional types of quarks and anti-quarks(sometimes referred to as ‘flavors').

To date, the existence of six quarks with their corresponding anti-quarks has been postulated.

They are divided into three 'generations'.

In addition to their electric charge, rest mass and intrinsic angular momentum (spin), each of these new quarks and its corresponding anti-quark is characterized by a previously unknown unique physical property.

These novel properties have been assigned whimsical names such as *strangeness*, *charm*, *bottomness* and *topness*.

The allowed values of these physical quantities are characterized by integer quantum numbers.

Table 5.5: The Parameters of the Six Quarks

	Quark Flavour	Unique Property	Rest Mass (MeV/c^2)[21]	Spin	Charge (e)
First Generation	Up		~300	½	$+\frac{2}{3}$
	Down		~300	½	$-\frac{1}{3}$
Second Generation	Strange	Strangeness	~500	½	$-\frac{1}{3}$
	Charm	Charm	~1500	½	$+\frac{2}{3}$
Third Generation	Bottom	Bottomness	~5000	½	$-\frac{1}{3}$
	Top	Topness	>50,000		

The existence of hadrons containing three identical quarks, all in the ground state and with their spins parallel, would appear to be forbidden by the exclusion principle; the occupation of the same orbital and spin state by three identical fermions is prohibited.

However, such a hadron, the relatively stable omega baryon, Ω^-, has been observed; it comprises three strange quarks with their spins parallel.

To get round this difficulty, it is simply postulated that the strange quark, as well as each of the other five quarks, can come in one of three different colors -red, blue and green- making a total of eighteen distinctive quarks each with its corresponding anti-quark.

It is postulated that each of the three strange quarks in the omega particle is of a different color.

Thus, they are not truly identical and so the exclusion principle is not violated.

This is the essence of the theory called *Quantum Chromodynamics*.

Color is an elusive quality of matter, but it lies at the heart of the quark hypothesis; it is the property that is the source of the strong force just as mass is the source of gravity and charge the source of electromagnetism.

And just as electromagnetic interactions can be interpreted in terms of an exchange of photons, so can both color and gravitational interactions.

The photon-like particles that mediate the former are the *gluons*; those that might be the mediators of gravity are called *gravitons*.

The basic interaction that holds hadrons together is the color force that acts between their constituent quarks by virtue of their color.

All normal hadrons are colorless; they contain counterbalancing amounts of color.

On this view, the strong force that acts between nucleons is a product of the color force.

It is analogous to the familiar van der Waals force that acts between atoms and molecules as a consequence of the electromagnetic force.

And like the van der Waals force, the strong force too falls off rapidly with distance.

However, here the similarity ends.

Unlike the electromagnetic force, the color force does not decrease with distance; it may even increase as the distance between the quarks grows.

It is this property that prevents the observation of free quarks.

The quarks can be pictured as joined to one another by an elastic string.

So long as the distance between them is less than the unstretched length of the string, they can move freely.

But should they move apart beyond this distance, the elasticity of the string brings them back together.

And should the string joining two quarks snap, the energy invested in the process appears as a pair of new quarks, one at each of the loose ends of the snapped string.

Thus, free unpaired quarks can never be produced.

Part 5.2 Nuclear Radiations

The true origin of the ionizing radiations first discovered by Becquerel remained an enigma until the nuclear model of the atom was postulated by Rutherford.

Subsequent research showed that these high energy radiations originated in the atomic nucleus and they were thus given the name nuclear radiations.

5.2.1 The Nature of the Nuclear Radiations

Initially. the radiations were classified on the basis of their penetrating power, i.e., in terms of the thickness of material required to attenuate a beam of the radiation.

In this way, three different types of radiation were identified: they were designated α , β and γ-rays.

Table 5.6: The penetrating power of nuclear radiations

Type	Penetrating Power
α-rays	Absorbed by a sheet of thin paper or cardboard.
β-rays	Absorbed by a 2mm sheet of lead.
γ-rays	Penetrate several centimetres of lead sheets before the beam is significantly attenuated.

More information about the radiations was obtained by passing them through electric and magnetic fields.

It was found that γ-rays are not deflected by either electric or magnetic fields; they carry no net charge.

However, both α-rays and β-rays are deflected, each in the opposite direction; α-rays are positively charged and β-rays negatively charged (Fig 5.8).

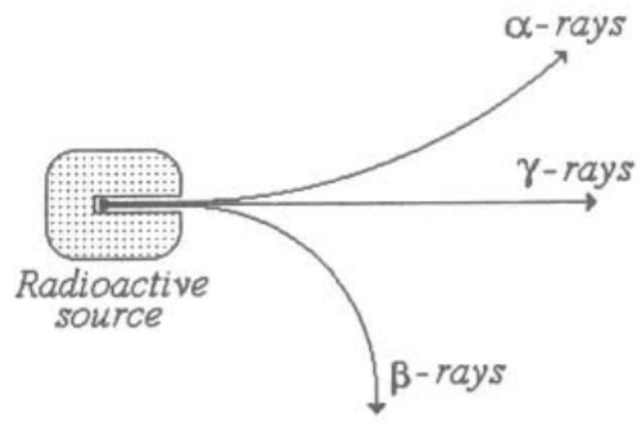


Fig 5.8 The deflection of α , β and γ-rays by a magnetic field whose direction is into the plane of the page.

The actual nature of the radiations emitted by radioactive substances was first elucidated by Rutherford in 1899, just three years after their discovery.

The ratio q/m between the charge and the mass of the β-rays was found to be the same as that of the electron.

This, together with their negative charge, sufficed to identify the β-rays emitted from the then known radioactive substances as electrons.

With the discovery and manufacture in the 1930's of radioactive isotopes of light elements, positively charged β-rays were detected.

These were found to be high energy positrons (anti-electrons).

The charge to mass ratio found for the α-rays corresponded to that of both the H_2^+ ion and the He_2^+ ion (helium nucleus).

In order to decide between these two alternatives, Rutherford collected α-rays in an evacuated discharge tube whose walls were too thick to allow their escape (Fig 5.9).

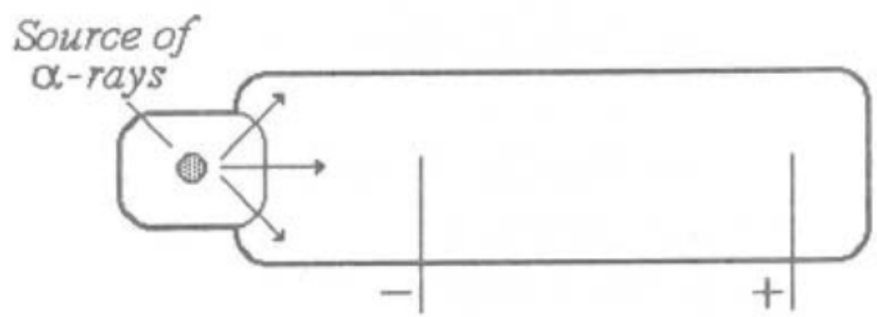


Fig 5.9 Rutherford's identification of α-rays. The α-rays cannot penetrate the thick walls of the compartment containing the radioactive source or those of the discharge tube. However, they can pass from the former to the latter through the thin wall between them. Sufficient time was allowed for a significant number of α particles to accumulate in the discharge tube. A high potential was then applied between the electrodes and light that exhibited the typical emission spectrum of helium filled the tube.

He found that when a high voltage was applied between the electrodes of the discharge tube, light that exhibited the typical emission spectrum of helium filled the tube.

This identified the α-rays as helium nuclei.

The more important features of the various nuclear radiations are summarized in Table 5.7.

Table 5.7 The properties of nuclear radiation

	α-rays	β-rays		γ-rays
		β^+	β^-	
Identity	Helium nuclei He^{2+}	Positrons	Electrons	Photons
Charge	$+2e$	$+e$	$-e$	0
Rest mass	4.0015u	0.00055u		0
Velocity	~0.06c	up to 0.98c		c
Energy	~6MeV	up to 1MeV		≥1MeV
Penetrating power	~10^5 ion-pairs per cm of air	~10^3 ion-pairs per cm of air		~10 ion-pairs per cm of air

With their identities established, the kinetic energies of the particles could now be determined; this required measuring their velocities.

Beams of α or β particles were passed through crossed electric and magnetic fields and the field strengths were adjusted until the particle beams passed through the crossed fields undeflected.

The particle velocities could then be calculated from the ratio, v = E/B, where E and B are the magnitudes of the respective fields.

It was found that the velocity of α particles is significantly lower than that of β particles.

However, both the helium nuclei and the electrons that, respectively, comprise the α and β-rays were found to possess considerable amounts of kinetic energy, of the order of millions of electron-volts.

5.2.2 Mechanisms of Nuclear Radiation Attenuation

As they pass through a material medium, the α or β particles emitted by radioactive sources repeatedly collide with and ionize the atoms and molecules comprising the medium.

In each such ionizing collision, a pair of charged particles is produced, an electron and a positive ion, and the incident α or β particle loses about 30eV of its kinetic energy, slowing it down accordingly.

In general, the penetrating power of ionizing radiation is inversely proportional to the number of such ion-pairs produced in each unit layer of material through which they pass.

The more ion-pairs produced in each unit layer, the thinner the thickness through which the ionizing α or β particles must travel before being stopped.

A typical α particle, with an initial energy of 3MeV, will produce about 100,000 ion-pairs before stopping completely.

Table 5.8 The stopping distances of α and β particles

Particle	Energy (MeV)	Stopping distance (cm) in: Air	Living Tissue	Aluminium
α	1	0.55	0.003	0.0003
	2	1.0	0.006	0.0006
	4	2.6	0.015	0.0015
β	0.01	12	0.015	0.004
	0.5	150	0.2	0.5
	1	420	0.5	0.15
	2	840	1.0	0.3

The relative penetrating powers of α and β radiation can be explained by the difference in the velocities of their component particles.

The massive and slower α particles (helium nuclei) are much more efficient ionizers than the lighter and faster β particles (electrons); they ionize almost every atom with which they collide losing some energy each time.

Consequently, α particles lose all their energy after passing through just a thin layer of matter (Fig 5.10).

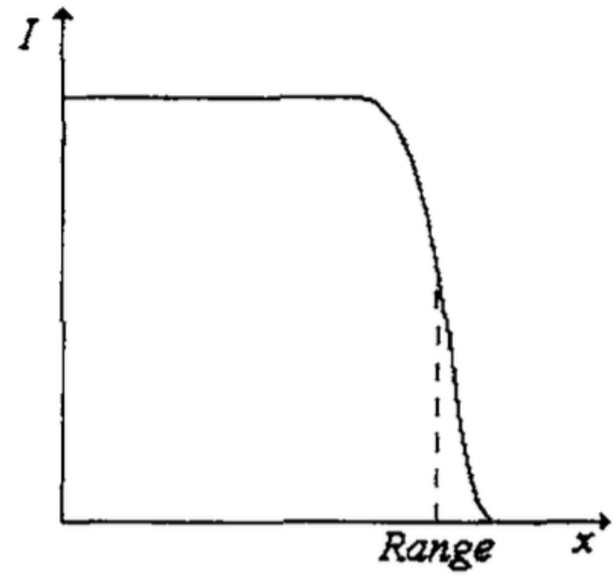


Fig 5.10 The intensity, I, of a beam of a radiation as a function of the thickness, x, of the material through which it passes. The intensity falls off sharply at a certain thickness called the range indicating that the particles all had the same initial energy. The range is approximately proportional to the density of the material.

However, the lighter and fleeter β particles are more easily deflected from their paths by the atoms and molecules of the material medium than are the heavier and more dogged α particles.

As a result, β particles do not ionize every atom with which they collide and so a thicker layer of material is required to stop them.

Moreover, the deceleration of the β electrons (or positrons) that accompanies these non-ionizing collisions generates X-rays (bremstrahlung).

The penetrating power of these X-rays exceeds that of the β particles and so even after all the β particles have been stopped by the material, these secondary X-rays may still be advancing through it.

The third type of nuclear radiation, the γ-rays, are high energy photons: $f > 10^{19}$Hz.

Like other forms of electromagnetic radiation, the intensity, I, of a beam of γ-rays is attenuated exponentially when passed through matter.

$$I(x) = I_0 e^{-\mu x} \tag{5.10}$$

In the case of γ-rays the coefficient μ is called the *linear attenuation coefficient or the macroscopic cross-section*.

Table 5.9 The thickness required to reduce the intensity of a beam of γ rays by a half.

Energy of the γ-ray photons (MeV)	Thickness of material (cm)	
	Living Tissue	Lead Metal
0.01	0.13	0.0008
0.1	4.0	0.01
1	9.8	0.9
5	23	1.5

The attenuation of high energy X-rays and γ-rays involves three different mechanisms.

1. <u>The photoelectric effect</u>

The photons in the incident high energy radiation release electrons from the atoms and molecules. The photons in the beam of radiation are annihilated, resulting in a corresponding reduction in its intensity.

2. <u>The Compton effect</u>

When a beam of high frequency electromagnetic radiation passes through a region containing free electrons an interaction takes place between the incident high frequency radiation and the free electrons in which energy is transferred from the radiation to the electrons. Relative to the high energies of the X-rays and γ-rays the outermost electrons of any atom are essentially free. As a result of the transfer of energy to the electrons the intensity of the radiation is reduced.

3. Pair production

When a positron and electron collide they self-destruct, releasing a burst of electromagnetic energy—photons. Conservation requires that the energy of these photons equal that of the annihilated particles. The rest energy of an electron or positron is 0.51MeV and so the energy of these photons must be at least 1.02MeV. The reverse process also occurs. When passing close to a nucleus, photons possessing an energy greater than 1.02MeV can transmute into an electron-positron pair (Fig 5.11). Charge, linear and angular momentum are all conserved in the process, that is usually referred to as *pair-production*.

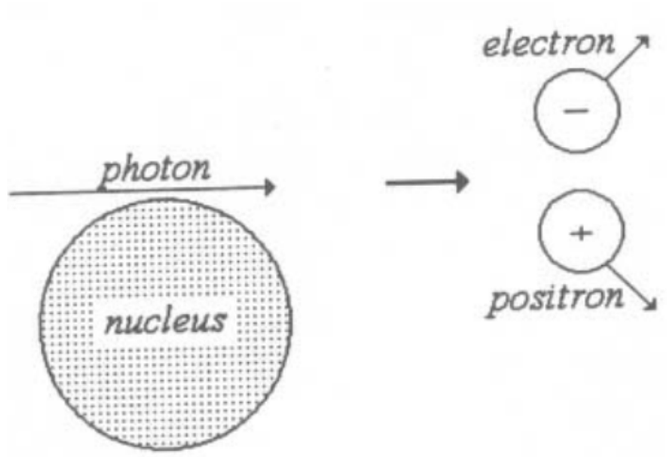


Fig 5.11 Pair production. The high energy photon is converted into an electron-positron pair. Linear momentum conservation requires that the electron and positron move away from each other.

The overall linear attenuation coefficient, μ, is the sum of the coefficients for the three mechanisms:

$$\mu = \mu_{Photoelectric} + \mu_{Compton} + \mu_{Pair-production} \tag{5.11}$$

For photons with an energy of between l MeV and 5 MeV the Compton effect is the primary absorption mechanism.

At higher energies, pair production becomes more prevalent (Fig 5.12).

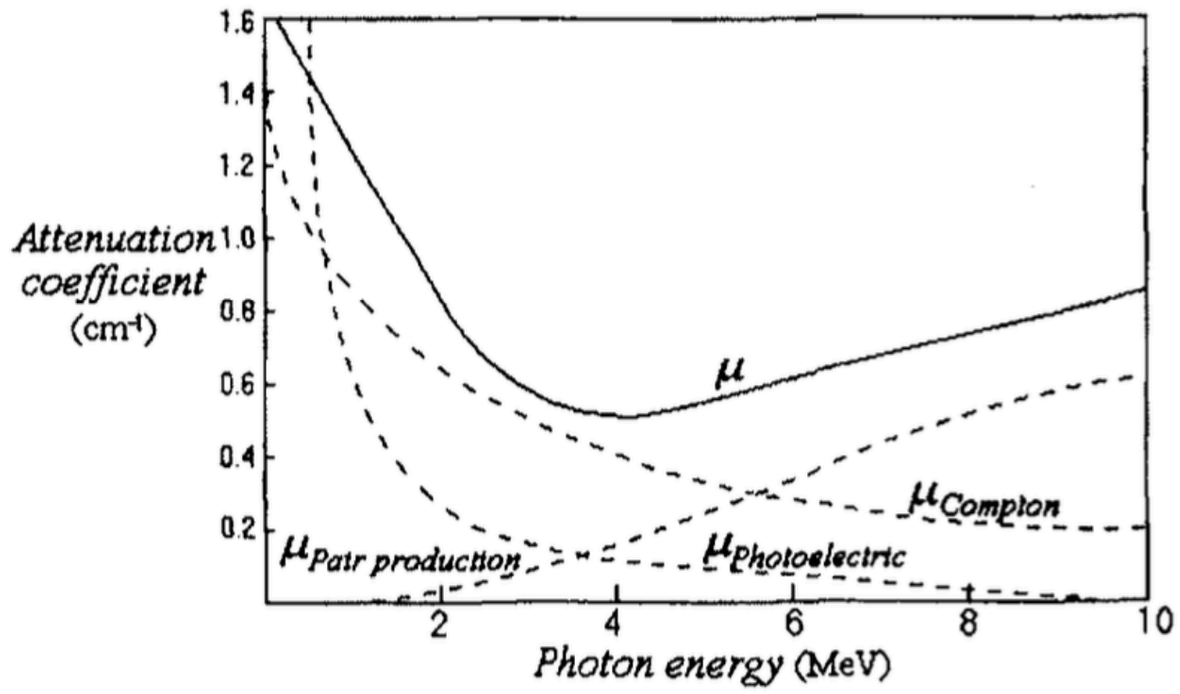


Fig 5.12 The linear attenuation coefficients of the three mechanisms of X-rayand γ-ray absorption as a function of the photon energy.

Although not emitted by naturally occurring sources of nuclear radiation, neutrons of both high and low energy are emitted from man-made sources such as atomic reactors.

The penetrating power of neutrons is very high and so atomic reactors must be encased in thick-walled containers.

Possessing no net charge, neutrons can quite easily reach into the nuclei of the material through which they are passing, disturbing their structure and thereby setting up secondary radiations.

Certain materials such as boron and cadmium absorb neutrons without setting up these secondary radiations and are therefore used as neutron absorbers.

5.2.3 Detectors of Ionizing Radiation

Ionizing radiation is detected by the ionization it causes.

Some of the most common devices used for their detection are briefly described here.

1. The cloud chamber

The ions produced as an ionising particle passes through clean air that is saturated with water or alcohol vapour, act as centres of condensation upon which tiny drops of water or alcohol are formed. These drops are formed all along the particle's path and so a visible track appears in the air. The cloud chamber is a device for producing such a saturated atmosphere and for viewing the tracks made by ionising particles. In the version invented by Wilson in 1912 the saturated atmosphere is produced by a sudden adiabatic expansion (Fig 5.13). The tracks formed by α-rays are thick and straight; those formed by β-rays are thin and curved. γ-rays eject electrons from air molecules which in turn produce a tangle of β-ray tracks.

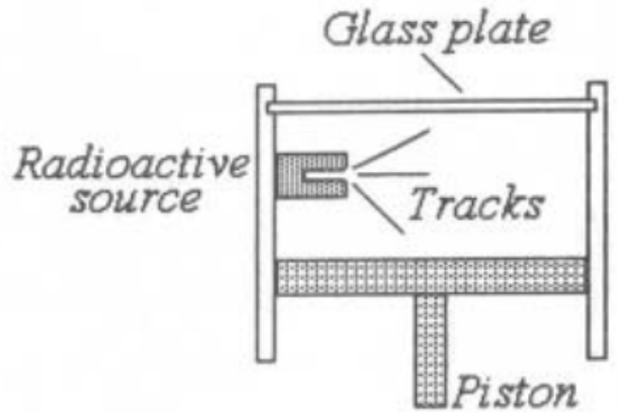


Fig 5.13 The Wilson cloud chamber. The air enclosed in the space above the piston is saturated with water vapor. Rapid withdrawal of the piston causes the air to cool and condensation results. The condensation takes place on the centers provided by the ions produced by the radiation emitted from the radioactive source and appears as tracks that can be viewed through the glass plate.

2. The bubble chamber

A liquid can be heated - superheated - above the temperature at which it normally boils if there are no 'centres' in the liquid bulk upon which bubbles can form. In the bubble chamber a liquid, typically liquid hydrogen, is heated under pressure. Suddenly reducing the pressure leaves the liquid in a superheated state. An ionising particle passing through the superheated liquid generates ions which act as ebullition centres. The tiny bubbles appear as a visible track in the liquid.

3. Nuclear emulsions

Ionising radiation can be detected by the fine tracks it leaves in sensitive photographic transparencies called *nuclear emulsions*. After developing, the tracks are revealed by examining the transparencies through a microscope. Typically, the tracks are just a few millimetres in length. Each type of ionising particles leaves a characteristic track by which it can be identified and its energy estimated. The emulsions are rugged and can be used where more delicate detectors would be inappropriate. For example, they can be taken aloft in unmanned balloons for the investigation of cosmic rays in the upper atmosphere.

4. The ionisation chamber - Geiger counter

The ionisation chamber is a gas-filled tube inside which a pair of electrodes have been installed. The gas, typically, a mixture of argon and a little ethyl alcohol vapour at low pressure (~0.1 atmospheres), is not an electric conductor under normal conditions. A potential difference is maintained between the electrodes (Fig 5.14).

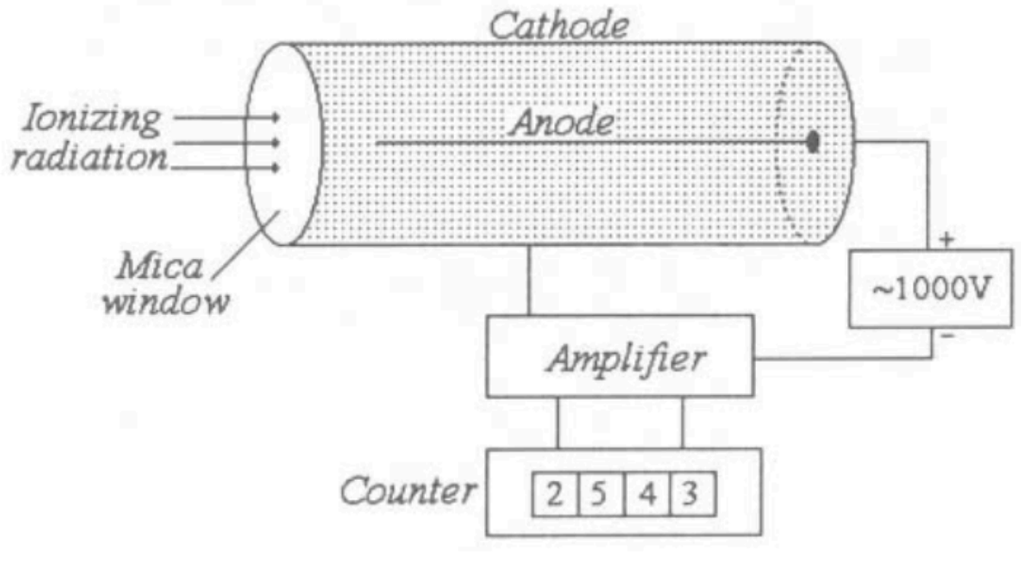


Fig 5.14 Geiger counter. Ionizing radiation enters the ionizing chamber through the thin mica window. The walls of the chamber are the cathode; the anode is a fine wire down its center.

The positive ions, Ar^+, and electrons liberated in the gas-filled tube when an ionising particle enters the tube are attracted to the cathode and anode, respectively. With a relatively high potential difference between the anode and cathode, ~1000V, the electron produced in the initial event collides with and ionises other argon atoms on its way to the anode, producing more argon ions and free electrons. The latter in turn generate still more ion-pairs inside the tube. A chain reaction takes place resulting in an avalanche of ionisation in the gas. The arrival of the free electrons at the anode is signalled by a detectable pulse of anode current. Since each pulse of anode current corresponds to the entry of an ionising particle into the tube, the number of particles can be determined by counting the pulses.

For the ionisation chamber to detect the next ionising particle, it must first be restored to its neutral state. This requires the neutralisation of the positive argon ions remaining in the gas. This is done by the ethyl alcohol molecules. Before reaching the cathode, where they might release electrons that would blur the pulses of anode current, the energetic argon ions collide with the alcohol molecules giving up their energy to them. Only positive alcohol ions reach the cathode where any excess energy they possess results in their decomposition rather than in the release of electrons.

Preparing the tube to receive the next ionising particle takes time, about a millisecond. During this 'dead time', as the argon atoms are being neutralised, the entry of an ionising particle would go undetected. Consequently, Geiger counters cannot be used for particle rates greater than a few hundred per second. They also give no indication of the energy of the ionising particles. They are employed mainly for counting β particles. They are inefficient counters of γ radiation or of high energy neutrons since not all of the photons or neutrons entering the tube induce the avalanche of ionisation in the gas.

5. Scintillation counter

In the famous experiment that demonstrated the existence of atomic nuclei, Geiger and Marsden observed the scattering of α particles by laboriously counting the light scintillations produced as each particle struck a zinc sulphide coated plate. In a modern scintillation counter this is done automatically by use of a *photomultiplier tube*.

When an ionising particle strikes a sodium iodide crystal a flash of light is produced. Some of this light enters the photomultiplier tube where it strikes the photocathode and releases a number of electrons from it by means of the photoelectric effect. These electrons are accelerated towards a second electrode that is held at a higher potential than the photocathode (Fig 5.15). Each electron that strikes this second electrode ejects a number of electrons from it, increasing the total number of electrons in the pulse. The shower of electrons is now accelerated towards a third electrode held at a still higher potential where the amplification process occurs once again. This is repeated a number of times until an electron pulse of significant strength is produced. Finally, the pulses are counted.

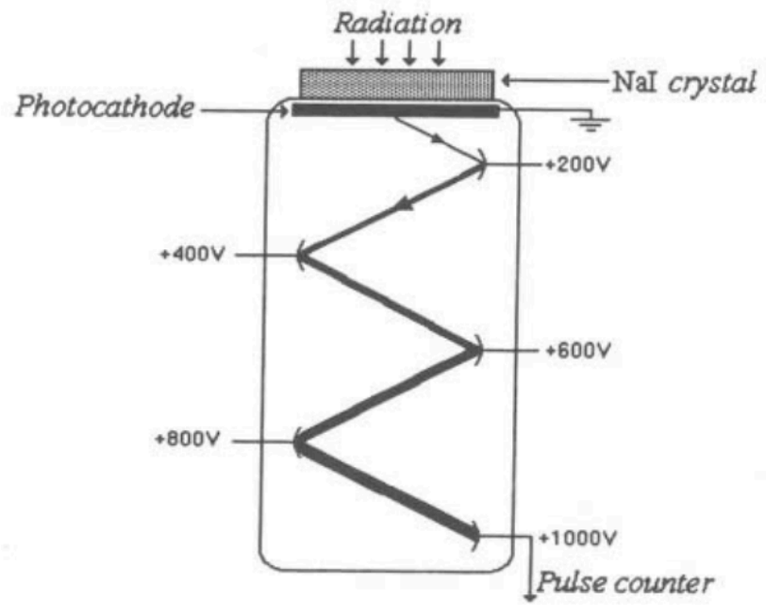


Fig 5.15 Scintillation counter. The entire device is kept in a light-proof container. It is also shielded from external magnetic fields. The number of electrons in the beam is amplified at each electrode.

Because the number of electrons emitted from the final electrode depends on the number incident on the first, each pulse recorded corresponds to a single ionising particle. Moreover, since the intensity of the light incident on the photocathode is determined by the energy of the original ionising particle, the strength of the pulse reflects the particle's energy. Thus, the scintillation counter not only counts the ionising particles but also measures their energy. The dead-time of a good scintillation counter is short, about a microsecond. It is also an efficient counter of both β and γ radiation.

By using light-emitting substances other than sodium iodide and by adjusting the potential differences between the electrodes, scintillation counters can be used to detect and measure ionising radiations over a wide range of energies and from different types of sources.

5.2.4 The Biological Effects of Nuclear Radiation

Since the major constituent of living tissue is water, the most probable initial result of the passage of an ionizing particle through living tissue is the ionization of a water molecule in reactions of the following type:

$$\begin{aligned} H_2O + \text{ionizing radiation} &\rightarrow H_2O^+ + e^- \\ H_2O + e^- &\rightarrow H_2O^- \end{aligned} \qquad (5.12)$$

The products of these reactions, H_2O+ and H_2O-, are very active species and can react with and cause damage to other molecules in the tissue.

In addition to the gross tissue damage they cause, these species can also induce genetic mutations.

Most of the biological damage that results from exposure to nuclear radiation arises from these secondary reactions.

The extent of the physical damage caused to living tissues by nuclear radiation is a function of four factors:

1. **the amount of radiation absorbed;**
2. **the nature of the ionising radiation;**
3. **the tissue or body part exposed to the radiation;**
4. **the rate at which radiation is absorbed.**

We will consider each of these factors, one by one.

1. The amount of radiation absorbed

The capacity of a source of radiation is often reported in a unit called the *Roentgen* whose symbol is R. For instance, the X-ray beam from a typical dental machine may have a rating of 250mR/s.[28] B t this says nothing about the radiation actually absorbed by the patient; the machine has the same rating whether anyone is sitting in the dentist's chair or not.

Living tissue is only affected by the radiation energy that it actually absorbs; it is unaffected by radiation that passes through it unabsorbed. The unit used to express the amount of radiation energy absorbed by living tissue is the *Gray* (Gy).

> An absorbed dose of 1 Gray equals the absorption of 1 Joule of radiation energy by 1 kg of tissue.

An alternative unit, used widely in the past, is the RAD (acronym of *Radiation Absorbed Dose*):

$$1\ \text{RAD} = 10^{-2}\text{Gy} = 10^{-2}\text{J/kg}$$

The dose received from a medical X-ray can range from 10μGy to 1000μGy.

2. The nature of the ionising radiation

The damage caused to living tissue varies according to the type of radiation to which it is exposed and how localised the energy absorption is. Densely ionising particles such as α particles and slow β particles, whose ranges in biological tissue are relatively short are more damaging than the high energy photons in X or γ rays. The relative biological effect of each type of ionising radiation is indicated by its *Quality Factor*—QF or *Relative Biological Effectiveness* (RBE).

Table 5.10: The quality factor (relative biological effectiveness) of various ionising radiations

Radiation	QF or RBE
γ or X-rays	1
Fast β particles	1
Slow β particles	2
Fast neutrons and protons	10
Slow neutrons	3 - 5
α particles	10 - 20

The effective absorbed dose, i.e., the biological effect of a dose of radiation, is the product of the absorbed dose (measured in Grays or RAD's) and the quality factor (or RBE) of the particular radiation. The effective absorbed dose is measured in units called *Sieverts* (Sv). For example, the biological effect of a dose of 10Gy of γ or X-rays

whose quality factor is equal to unity is $10 \cdot 1 = 10Sv$. A dose of 5Gy of slow β particles or of 0.5Gy of α particles would have the same biological effect.

An alternative unit, frequently used in the past to measure the biological effect of radiation, is the REM (acronym of *RAD Equivalent Mammal*[29]):

$$100 \text{ REM} = 1Sv$$

3. The tissue or body part exposed to the radiation

Because of their greater density, bones absorb relatively more radiation than soft tissue. The cells most sensitive to radiation damage are those which have the highest rate of division, and are the least differentiated. Cancer cells exhibit these characteristics and so are more susceptible to radiation damage than other cells, This is the basis of the treatment of cancer by high energy radiations. On the other hand, these characteristics are also typical of zygotes and young embryos and so great care must be taken in the irradiation of women of child-bearing age. A dose of 40Sv to the gonads produces sterility in men; a dose of just 10Sv renders women sterile.

Red blood cells are more resistant to radiation damage than white blood cells. Changes in the ratio between the number of red and white cells in the blood can be used to measure the doses to which an individual has been exposed. Whole body doses as low as 0.25Sv can affect the ratio between the cells. Apart from the gross tissue damage inflicted by an exposure to ionising radiation, mutations in the organism's genetic material may also occur. Whether genetic damage occurs only above a certain minimum threshold of radiation or at any level, however low, is still unresolved.

3. The tissue or body part exposed to the radiation

Because of their greater density, bones absorb relatively more radiation than soft tissue. The cells most sensitive to radiation damage are those which have the highest rate of division, and are the least differentiated. Cancer cells exhibit these characteristics and so are more susceptible to radiation damage than other cells, This is the basis of the treatment of cancer by high energy radiations. On the other hand, these characteristics are also typical of zygotes and young embryos and so great care must be taken in the irradiation of women of child-bearing age. A dose of 40Sv to the gonads produces sterility in men; a dose of just 10Sv renders women sterile.

Red blood cells are more resistant to radiation damage than white blood cells. Changes in the ratio between the number of red and white cells in the blood can be used to measure the doses to which an individual has been exposed. Whole body doses as low as 0.25Sv can affect the ratio between the cells. Apart from the gross tissue damage inflicted by an exposure to ionising radiation, mutations in the organism's genetic material may also occur. Whether genetic damage occurs only above a certain minimum threshold of radiation or at any level, however low, is still unresolved.

Ingesting radioactive materials is particularly dangerous. The fall-out from the explosion of a typical 1 megaton nuclear weapon contains about 0.5kg of the relatively long-lived radioactive strontium isotope Sr–90.$^{3-}$ Chemically, strontium is similar to calcium and it was found during the 1950's that radioactive Sr–90, released by the atmospheric testing of nuclear weapons, was accumulating in the bones of children fed the milk of cows that had grazed in regions where the fall-out came to earth. The long-term implications of this were considered so dangerous that the nuclear powers agreed to cease the atmospheric testing of nuclear weapons. After the Chernobyl disaster in 1986, the meat of lambs that had grazed in wet areas of the UK was taken off the market, for fear that they had become contaminated by the radioactive fall-out the rain had brought down in the days and weeks after the explosion.

4. The rate at which the radiation is absorbed

Whereas a single dose of 10Sv can be lethal, certain cancer patients may receive daily doses of 1Sv for a period of a fortnight, making a total of 14Sv for the period. Evidently, the rate at which the radiation is received is an important factor in determining its biological effect.

Skin reddening, *erythema*, is one of the manifestations of tissue damage caused by high-energy radiation and it has been used to estimate the effects of radiation. Table 5.11 shows the effect of different radiation rates on skin. Clearly, substantial doses can be tolerated if they are given slowly and the tissue is allowed time to repair.

Table 5.11: Doses of X-ray irradiation that produce the same degree of erythema

Total amount of radiation (Gy)	Irradiation time (mins)	Rate of radiation (Gy/min)
5.0	1	5
7.8	15	0.50
13.0	260	0.05
22.5	4500	0.005

The whole body dose that kills 50% of a population of organisms is designated 'LD_{50}'. A burst of 30Sv of X-rays will kill 50% of a newt population; the LD_{50} for guinea pigs is just 2.5Sv. On the basis of the mortality rates recorded in Hiroshima and Nagasaki after their devastation by nuclear weapons in 1945, it has been estimated that the human LD_{50} is about 6.5Sv. Protecting the more sensitive organs such as the spleen increases the chances of survival.

Table 5.12 Annual individual exposure to ionising radiation in the UK

Radiation Source	Dose (μSv)	
Natural Sources		
Cosmic rays	300	
Terrestrial sources	400	
Radioactive isotopes in the human body	370	
Radon	700	
Thorium	100	
		1870
Artificial Sources		
Medical	250	
Fall-out from weapon testing	10	
High-altitude flying, watching TV etc.	10	
Occupational	8	
Nuclear industry	2	
		280
Total		2150

The whole body dose that kills 50% of a population of organisms is designated 'LD_{50}'. A burst of 30Sv of X-rays will kill 50% of a newt population; the LD_{50} for guinea pigs is just 2.5Sv. On the basis of the mortality rates recorded in Hiroshima and Nagasaki after their devastation by nuclear weapons in 1945, it has been estimated that the human LD_{50} is about 6.5Sv. Protecting the more sensitive organs such as the spleen increases the chances of survival.

We are all exposed to natural sources of ionising radiation such as cosmic rays and the naturally occurring radio-isotopes such as the radon gas emitted by rocks and the ^{40}K in our own bodies (Table 5.12). Thus, total avoidance of high-energy radiation is impossible. In considering the risks implicit in man-made radiation to which we are exposed, the undoubted benefits such as those of medical X-rays must also be taken into account. Whilst there is probably no completely safe dose of ionising radiation, there is also no entirely safe way of crossing the road; caution is required in both cases. Clearly, the experience of the century since the discovery of X-rays has shown that in small and intermittent doses, ionising radiation is certainly not lethal. Whether these doses are as safe genetically in the long term is impossible to gauge at present.

The United States Environmental Protection Agency has set an upper limit of 0.0017Sv for the annual permitted whole-body dose, over and above the ionising radiation to which we are exposed from natural and medical sources. An overall annual upper limit from all sources of 0.005Sv is recommended by most authorities, though workers in occupations that involve ionising radiations appear to tolerate annual exposures as high as 0.05Sv without harm.

Part 5.3 Radioactivity

Of the more than 1400 known atomic nuclei, only about 280 are stable.

All the stable nuclei occur naturally.

Some of the unstable ones, particularly those of the heavy elements, are also found naturally.

All the other unstable nuclei are manufactured from naturally-occurring nuclei by nuclear reactions.

These are usually referred to as artificial nuclei and the atoms they comprise as artificial isotopes.

In general, unstable nuclei disintegrate spontaneously by the emission of nuclear radiation they are *radioactive*.

When describing nuclear decays, the disintegrating nucleus is usually referred to as the parent nucleus and the nucleus remaining after the event as the *daughter nucleus*.

The total rest mass of the daughter nucleus and of the nuclear radiation released in a radioactive disintegration, $m_{Daughter} + m_{Radiation}$, is always less than that of the parent nucleus, m_{Parent}.

The mass difference,

$$\Delta m = (m_{Daughter} + m_{Radiation}) - m_{Parent}$$

appears as the disintegration energy, $Q = \Delta mc^2$, liberated in the process.

The mass number, A, of the daughter nucleus produced when an α particle is emitted from a radioactive nucleus is always four units less than that of the parent nucleus and its atomic number, Z, is always two less.

For example:

$$\underset{\substack{\textit{Parent}\\\textit{Nucleus}}}{{}^{226}_{88}\mathrm{Ra}} \rightarrow \underset{\substack{\textit{Daughter}\\\textit{Nucleus}}}{{}^{222}_{86}\mathrm{Rn}} + \underset{\substack{\textit{Alpha}\\\textit{Particle}}}{{}^{2}_{4}\mathrm{He}} + \underset{\substack{\textit{Disintegration}\\\textit{Energy}}}{4.78\mathrm{MeV}} \tag{5.13}$$

The emission of a β particle, either an electron, β-, or a positron, β+, changes the atomic number of the nucleus without affecting its mass number.

In β- emission, a neutron in the parent nucleus becomes a proton in the daughter nucleus; this raises the atomic number by one.

$$\underset{\substack{\textit{Parent}\\\textit{Nucleus}}}{{}^{214}_{83}\mathrm{Bi}} \rightarrow \underset{\substack{\textit{Daughter}\\\textit{Nucleus}}}{{}^{214}_{84}\mathrm{Po}} + \underset{\substack{\textit{Beta}\\\textit{Particle}}}{e^{-}} + \underset{\substack{\textit{Disintegration}\\\textit{Energy}}}{3.26\mathrm{MeV}} \tag{5.14}$$

In β+ emission, a proton in the parent nucleus becomes a neutron in the daughter nucleus; this reduces the atomic number by one.

$$\underset{\substack{\textit{Parent}\\\textit{Nucleus}}}{{}^{25}_{13}\mathrm{Al}} \rightarrow \underset{\substack{\textit{Daughter}\\\textit{Nucleus}}}{{}^{25}_{12}\mathrm{Mg}} + \underset{\substack{\textit{Beta}\\\textit{Particle}}}{e^{+}} + \underset{\substack{\textit{Disintegration}\\\textit{Energy}}}{3.26\mathrm{MeV}} \tag{5.15}$$

After an α or β emission, the daughter nucleus is often left in an excited energy state.

It subsequently reaches its ground state by emitting high-energy photons, γ-rays.

$$
\begin{array}{lcccc}
{}^{12}_{5}\mathrm{B} & \rightarrow & {}^{12}_{6}\mathrm{C}^{*} & + & e^{-} \\
\textit{Parent Nucleus} & & \textit{Excited Daughter Nucleus} & & \textit{Beta Particle} \\
{}^{12}_{6}\mathrm{C}^{*} & \rightarrow & {}^{12}_{6}\mathrm{C} & + & \gamma \\
\textit{Excited Daughter Nucleus} & & \textit{Ground - state Daughter Nucleus} & & \textit{Gamma Particle - Photon}
\end{array}
\qquad (5.17)
$$

The emission of a high-energy photon affects neither the mass number nor the atomic number of the nucleus.

However, it does reduce its rest-mass by an amount, $\Delta m = hf/c^2$, equivalent to that of the photon's energy.

A typical nuclear energy level diagram illustrating this process, is shown in Fig 5.16.

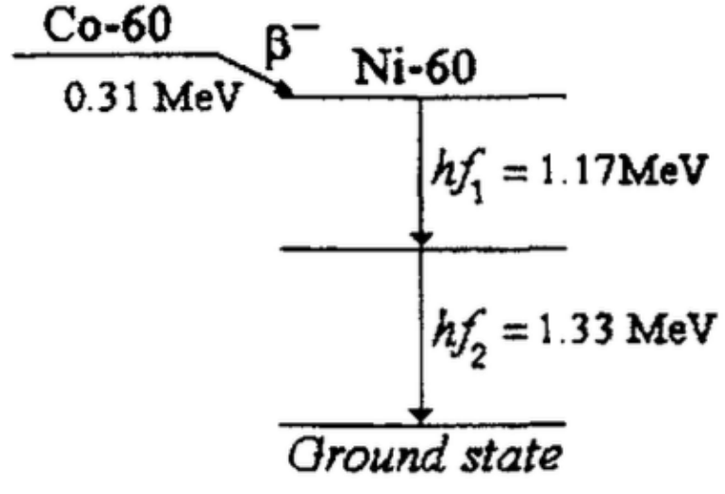


Fig 5.16 The Ni-60 nucleus produced from the emission of a α particle from Co-60 is in an excited nuclear energy state. It falls to its ground state in two stages, emitting first a photon of energy 1.17MeV and then another of I.33MeV

Almost all nuclei with an atomic number greater than 82 - the atomic number of the element lead, Pb - are radioactive.

Since the only change in the mass number of a nucleus that can occur as a result of radioactive decay (α emission) is -4, all these unstable heavy nuclei can be fitted into one of four decay sequences or series, each characterized by the mass number, A, of its component nuclei: 4n, 4n + 1, 4n + 2 and 4n + 3, where n is an integer between 51 and 59.

These are known, respectively, as the thorium, neptunium, uranium and actinium series:

Table 5.13 The four radioactive decay series

Series	Parent Nucleus	Mass numbers
1. Thorium series	$^{232}_{90}$Th	$A = 4n$
2. Neptunium series	$^{237}_{93}$Np	$A = 4n + 1$
3. Uranium series	$^{238}_{92}$U	$A = 4n + 2$
4. Actinium series	$^{235}_{92}$Ac	$A = 4n + 3$

A nuclear proton can also be converted into a nuclear neutron by the capture of an electron.

This process competes with β+ emission.

$$\underset{\substack{\textit{Parent}\\\textit{Nucleus}}}{{}^{15}_{8}\mathrm{O}} + \underset{\textit{Electron}}{e^{-}} \rightarrow \underset{\substack{\textit{Daughter}\\\textit{Nucleus}}}{{}^{15}_{7}\mathrm{N}} + \underset{\substack{\textit{Disintegration}\\\textit{Energy}}}{2.75\mathrm{MeV}} \qquad (5.16)$$

The daughter nucleus produced from the decay of a radioactive nucleus is often itself radioactive.

This can result in a sequence of decays. For example:

$$\begin{array}{ccccccccc} {}^{238}_{92}\mathrm{U} & \rightarrow & {}^{234}_{90}\mathrm{Th} & \rightarrow & {}^{234}_{91}\mathrm{Pa} & \rightarrow & {}^{234}_{92}\mathrm{U} & \rightarrow & {}^{230}_{90}\mathrm{Th} \rightarrow \text{etc.} \\ & & +\alpha & & +\beta^{-} & & +\beta^{-} & & +\alpha \end{array}$$

All the intermediate nuclei of a decay series are less stable than the parent nucleus. The thorium, uranium and actinium series, which occur naturally, all end with a stable isotope of the element lead (Fig 5.17).

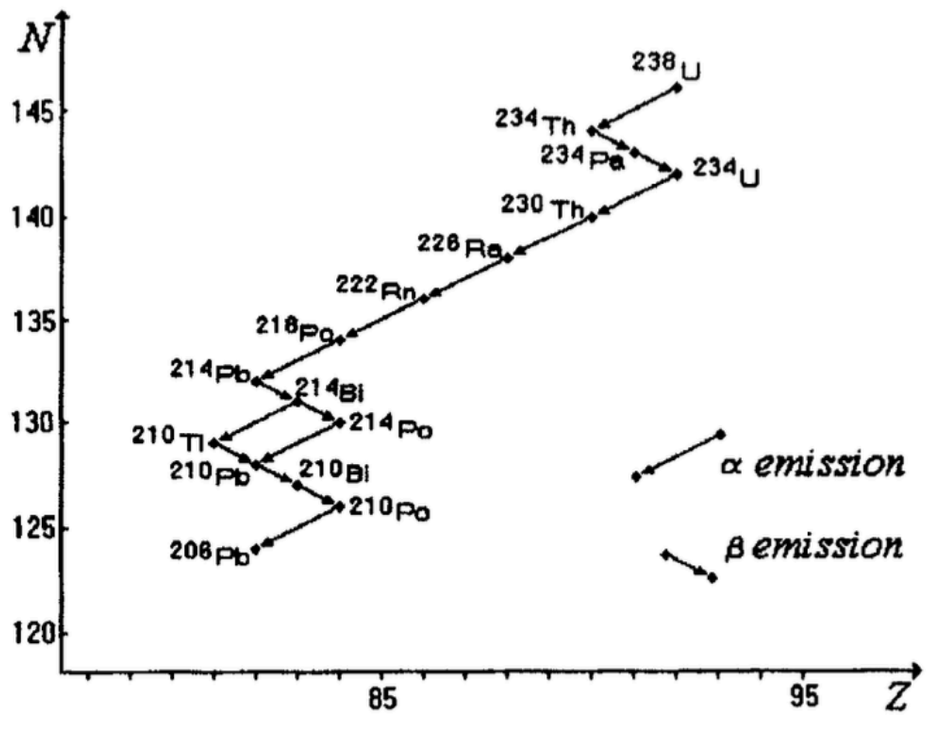


Fig 5.17 The uranium series. The Bi-214 nucleus can decay either by α emission or by β- emission. However, both routes lead ultimately to the Pb-210 nucleus.

The parent nucleus of the neptunium series is not found in nature, though it can be manufactured.

Compared to the parent nuclei of the other series it is very unstable.

Thus, within the estimated age of the solar system, it and any of the intermediate nuclei of its series that might once have existed have all decayed into the stable nucleus, Bi-209, with which the series ends.

5.3.2 The Kinetics of Radioactive Disintegration

The disintegration of a radioactive nucleus is a random process; there is no way of predicting precisely when a particular unstable nucleus will disintegrate.

'Old' radioactive nuclei that have been around for a long time are no more or less likely to disintegrate in the next minute than freshly produced but otherwise identical 'young' nuclei.

Similarly, the probability that a nucleus will disintegrate is unaffected by its environment, be it hot or cold, high pressure or low pressure, or whether it is a component of one chemical compound or another.

It follows, that the rate at which a sample of radioactive matter decays and emits nuclear radiation, depends only on the number of radioactive nuclei it contains.

The rate of decay, dN/dt , of a sample containing N radioactive nuclei is given by:

$$\frac{dN}{dt} = -\lambda N \tag{5.18}$$

The rate of decay is negative because the number of radioactive nuclei decreases with time.

The parameter, λ, called the *decay constant*, is a characteristic of the unstable nuclei.

The absolute value, ldN/dtl, of its rate of decay is called the sample's *activity*.

The units in which activities are expressed are the *Becquerel* whose symbol is Bq or the *Curie* which is designated Ci.

1Bq is an activity of one disintegration per second: $1\text{Ci} = 37 \cdot 10^9\ \text{Bq}$

Radioactive sources used in cancer treatment, such as those containing the isotope Co-60, have activities of about 10^{14}Bq.

In the explosion of the nuclear reactor and power station at Chernobyl in 1986, radioactive materials with an activity of at least 90-106Ci were liberated into the atmosphere.

This is equivalent to at least 6750Bq on each square meter of the earth's surface, an amount of radioactive fall-out comparable to that it is estimated would be released in a medium scale nuclear strike.

Radioactive sources used in student laboratories have activities of 105Bq or less.

Antique luminous watches whose numerals are painted with uranium-based paints have activities of about 20,000Bq.

The solution of equation (5.18) is:

$$N_t = N_0 e^{-\lambda t} \tag{5.19}$$

where N_0 is the number of unstable nuclei in the sample at t = 0.

This exponential law of radioactive disintegration was first discovered by Rutherford and Soddy in 1902 (Fig 5.18).

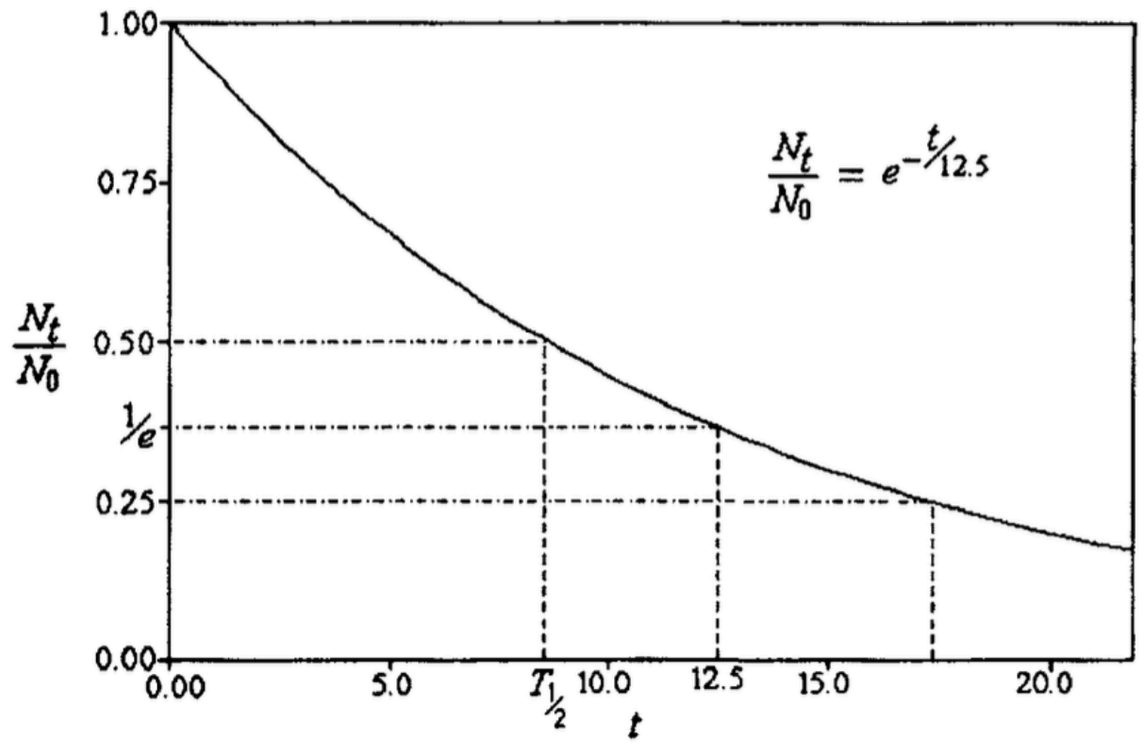


Fig 5.18 The Rutherford-Soddy law of radioactive decay. The graph shows the quotient N_t/N_0 of the number of unstable nuclei in a sample as a function of the time, t, where τ = 12.5 time units. A graph of the sample's activity, |dN/dt|, as a function of time would have exactly the same form.

The disintegration law can also be expressed in terms of the parameter $\tau = 1/\lambda$, the average lifetime of the unstable nucleus.

$$N_t = N_0 e^{-t/\tau} \tag{5.20}$$

The most commonly quoted parameter of an unstable nucleus is its *half-life*, $T_{1/2}$.

It is the length of time after which half the nuclei of its type will have disintegrated in any given sample of them.

In a sample containing N_0 unstable nuclei at any given moment, it is the time after which $N_t = N_0/2$.

Substitution in equation (5.19) gives:

$$N_{0/2} = N_0 e^{-\lambda T_{1/2}} \tag{5.21}$$

which reduces to

$$2 = e^{\lambda T_{1/2}} \tag{5.22}$$

or

$$T_{1/2} = \ln 2/\lambda = 0.693/\lambda \tag{5.23}$$

5.3.3 Age Determination with Radio-isotopes

Because the rate at which radioactive nuclei disintegrate is unaffected by their environment, it can be used to estimate the age of any material sample or object which contains a radioactive isotope.

In addition to a knowledge of the radio-isotope's half-life and of its present abundance and that of its decay products in the sample, this dating method also requires making certain assumptions about their original abundance.

Thus, for example, the age of the earth's crust can be estimated from the ratio between the amounts of U-238 and Pb-206 found in geological specimens.

These two isotopes are respectively the first and last members of the uranium decay series, all the intermediate members of which are so unstable that the decay of the parent radioactive uranium can be considered to lead directly to the stable lead isotope.

The assumption made is that all the Pb-206 nuclei found in the specimen to-day were originally U-238 nuclei, i.e., that at the crust's formation the specimen contained no Pb-206 nuclei.

If no other lead isotopes are found in the specimen, this is a reasonable assumption.

Typically, geological samples contain equal numbers of U-238 nuclei and Pb-206 nuclei, i.e., half of the uranium nuclei they originally contained have transmuted into lead.

Thus the age of the earth's crust approximately equals the half-life of U-238, viz. $4.5 10^9$ years.

The carbon radio-isotope C-14 is frequently used for dating archaeological finds.

This isotope is formed in the upper atmosphere from the collision of high-energy neutrons with nitrogen atoms in the air.

$$^{14}_{7}\mathrm{N} + n \rightarrow {}^{14}_{6}\mathrm{C} + p$$

The high-energy neutrons are a constituent of the shower of particles produced when cosmic ray protons enter the atmosphere and collide with the atoms in the air molecules.

The half-life of C-14 is 5730 years.

The stable C-12 isotope is by far the most common carbon isotope in atmospheric carbon dioxide.

On average just one out of every $1.3 10^{12}$ carbon atoms in the atmosphere is a C-14 atom.

As a result of photosynthesis, the same ratio is found in green plants and subsequently in the food chain of living organisms.

However, once an organism dies it ceases to absorb any more C-14.

Assuming that the incidence of both carbon isotopes in the atmosphere has remained unchanged over the period, the age of the remains of a once living object can be determined by comparing the ratio of C-14 to C-12 atoms found in its remains to that in the atmosphere.

The method enables datings to be made up to about 20,000 years ago with an accuracy of about ±100 years.

5.3.4 Uses of Radio-isotopes

Apart from their usefulness in dating, radio-isotopes are widely employed both in research and industry.

Because (i) radio-isotopes are chemically identical to the stable isotopes of the same element and (ii) the radiation they emit can be detected with relative ease, they are most frequently used as tracers.

For instance, the transport of sodium and potassium ions across cell membranes can be followed with the radio-isotopes Na-22 and K-42. Similarly, binding to membranes, proteins and DNA molecules can be tracked with radio-tracers such as Ca-45.

The take up of fertilizers by plants is studied by labelling the fertilizers with a minute amount of the radio-isotope P-32 and the activity of the thyroid gland can be examined by giving the patient a dose of the radio-isotope I-131.

Table 5.14: The half-lives and radiations emitted from a number of radioactive isotopes

Isotope	Half-life	Radiation	Energy (MeV)
H–3	12 years	β^-	0.018
C–14	5730 years	β^-	0.160
Na–22	2.6 years	β^+	0.51
		γ	1.28
P–32	14 days	β^-	1.7
S–35	87 days	β^-	0.17
K–40	$1.3 \cdot 10^9$ years	β^-	1.3
		γ	1.5
K–42	12 hours	β^-	2.0
		γ	3.6
I–131	8 days	β^-	0.610
		γ	0.360

The high-energy penetrating radiation emitted by radio-isotopes has been used for the treatment of deep tumors, for the continuous measurement of the thickness of metal sheeting and the sterilization of heat-sensitive disposable medical equipment such as plastic syringes.

In the technique called *Positron Emission Tomography* - *PET*, real-time images of the functioning of body organs including the brain have been obtained.

The person or animal under examination receives an injection of a short-lived $\beta+$ emitting radio-isotope such as ^{15}O.

The moment they are emitted, the positrons released by the isotope react the with an electron in the surrounding matter producing two γ photons.

These photons are detected and located by a bank of scintillation counters.

Images of the motion of the radio-isotope can be obtained by feeding the locations into a computer.

The technique has been used to detect changes in blood flow in the heart and brain as a conscious subject performs various tasks or is stimulated.

5.3.5 The Factors Affecting Nuclear Stability

Although no comprehensive nuclear theory has yet been developed, certain factors that appear to affect nuclear stability can be clearly identified.

1. **The ratio between the number of neutrons and protons**

A plot of the number of neutrons, N, in the 280 stable nuclei versus the number of protons, Z, suggests that they constitute a band of stability (Fig 5.19). Apparently, the ratio, N/Z, is a factor in their stability. Whereas light nuclei ($A < 20$) contain approximately equal numbers of neutrons and protons, the ratio between the number of neutrons and protons rises as the total number of nucleons in the nucleus increases.

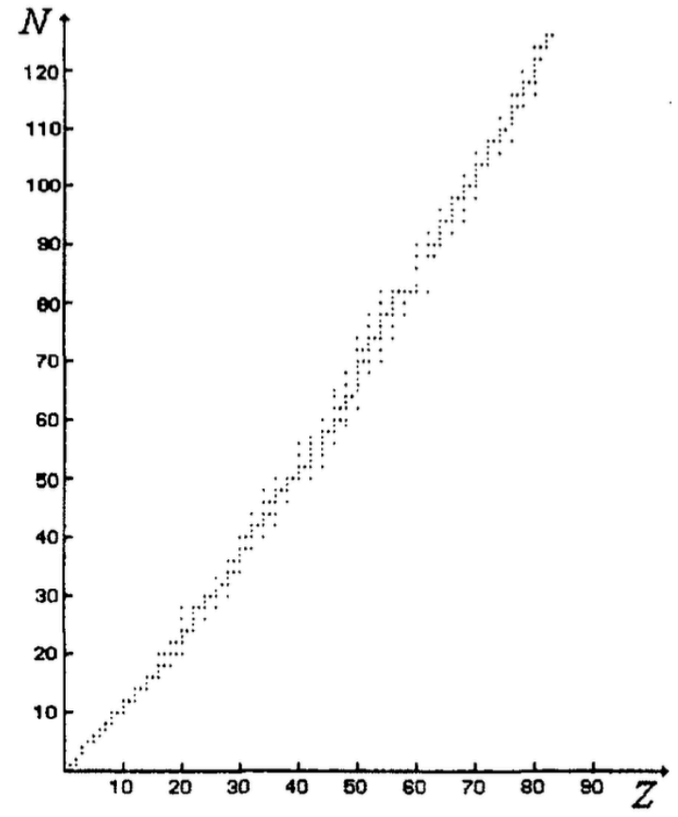


Fig 5.19 The band of stable nuclei. There are no stable nuclei of atomic number 43 or 61; with 19, 21, 35, 39, 45, 61, 89, 115, or 123 neutrons; or containing a total of 5 or 8 nucleons. All nuclei with atomic number greater than 83, or those containing more than 126 neutrons or 209 nucleons, are unstable (radioactive).

2. The size of the nucleus

All nuclei containing more than 210 nucleons are unstable. Evidently, size too affects nuclear stability. This is attributed to the very short-range over which the cohesive nuclear force acts. As the size of the nucleus increases, the long-range electrostatic repulsive interaction between its protons becomes more dominant and even the substantial neutron excess found in these large nuclei cannot overcome it.

3. The pairing of the nucleons

Except for just five nuclei, ${}^{2}_{1}H$, ${}^{6}_{3}Li$, ${}^{10}_{5}Be$, ${}^{14}_{7}N$ and ${}^{180}_{73}Ta$, all the stable nuclei contain an even number of protons or an even number of neutrons. Sixty percent contain even numbers of both protons and neutrons. Thus the pairing of its nucleons appears to be a factor in the stability of a nucleus. This is consistent with the presence of nuclear energy levels each of which, in compliance with the exclusion principle, can be occupied by no more than two protons or two neutrons.

4. Magic numbers of nucleons

Nuclei having 2, 8, 20, 28, 50, 82 or 126 protons or neutrons are more abundant in nature than those having similar mass numbers but different compositions. Nuclei with these magic numbers of nucleons are also relatively more stable. This suggests that nucleonic configuration is also a stability factor.

On the basis of these four factors, the particular type of radiation emitted by different unstable nuclei, in their search for stability, can be explained.

4. Magic numbers of nucleons

Nuclei having 2, 8, 20, 28, 50, 82 or 126 protons or neutrons are more abundant in nature than those having similar mass numbers but different compositions. Nuclei with these magic numbers of nucleons are also relatively more stable. This suggests that nucleonic configuration is also a stability factor.

On the basis of these four factors, the particular type of radiation emitted by different unstable nuclei, in their search for stability, can be explained.

a. Nuclei having a low neutron/proton ratio

If its instability arises from too low a neutron/proton ratio, a radioactive nucleus can improve the situation by β^+ emission, i.e., by emitting a positron or capturing an electron. In the process, one of its protons is converted into a neutron, thereby raising the ratio. Typically, such radioactive nuclei lie to the right of the stability band. If, on the other hand, its instability arises from too large a neutron/proton ratio, stability will be gained by β^- emission in which a neutron is converted into a proton. Typically, such nuclei lie above and to the left of the band.

b. Nuclei with unstable configurations

The instability of the radioactive nuclei that occupy the empty spaces in the stability band is associated with their nucleonic configurations. In many instances, these nuclei embody unpaired nucleons. In such cases, the stability of the nuclei can be improved by β radiation. For example, if an unpaired neutron in the nucleus decays by β^- emission into a proton, and the proton into which it transmutes pairs up with another previously unpaired proton, the total energy of the nucleus will generally be reduced. Overall, two previously unpaired nucleons will have become paired.

c. Nuclei containing more than 200 nucleons

Nuclei, that are unstable because they contain too many nucleons, cannot gain stability by β emission since this leaves their mass number unchanged. In principle, such nuclei could reduce their nucleon excess by the emission of free nucleons (protons or neutrons), deuterons ($^2H^+$) or helium nuclei ($^3He^{++}$ or $^4He^{++}$). However, as the calculations in Example 5.8 show, the emission of an α particle ($^4He^{++}$) is the only one of these disintegration modes that is energetically possible for heavy nuclei. As a result, most unstable heavy nuclei with a mass number $A > 200$, are α emitters. The key to this is the considerable binding energy, 28MeV, that the $^4He^{++}$ nucleus can make available to the disintegrating nucleus. None of the other disintegration modes have anything like this amount of energy available to them (Table 5.15).

Table 5.15:The binding energies of helium nuclei, deuterons and a free nucleons.

Particle	Binding Energy
α particle ($^4He^{++}$)	28 MeV
$^3He^{++}$	7 MeV
deuteron	2 MeV
nucleon	0 MeV

Having identified the factors that determine nuclear stability and the choice of the decay mode by which unstable nuclei can attain stability, we must now address a more fundamental question:

Why do unstable nuclei still exist?

Why haven't all the naturally occurring radioactive nuclei decayed long ago into stable nuclei?

If these entities are inherently unstable, what mechanism prevents them from instantly disintegrating?

This is a key question that must be addressed by the theories postulated to explain α and β decay.

5.3.6 The Mechanism of a Decay

Inside the parent nucleus, the potential energy of the component nucleons of an α particle -the two protons and two neutrons- must be negative since they are bound by the strong force to the rest of the nucleus.

However, the moment it finds itself outside the nucleus, the α particle has a positive potential energy due to its electrostatic repulsive interaction with the daughter nucleus.

Consider, for example, the emission of an a particle from the $^{226}_{88}\mathrm{Ra}$ nucleus.

Assuming the nucleus to be a sphere of radius R = $1.2(222)^{1/3} = 7.3\text{fm}$, the electrostatic potential energy, U_E, of the α particle on the surface of the daughter nucleus, $^{226}_{88}\mathrm{Rn}$, is, according to Coulomb's law:

$$
\begin{aligned}
U_E &= \frac{kq_\alpha q_{\mathrm{Rn}-222}}{R} \\
&= \frac{9\cdot 10^9 \cdot 2 \cdot 86 \cdot \left(1.6\cdot 10^{-19}\right)^2}{7.3\cdot 10^{-15}\cdot 10^6 \cdot 1.6\cdot 10^{-19}} \\
&= 34\text{MeV}
\end{aligned}
\tag{5.24}
$$

The energy, ~4.6MeV, made available to the a particle by the decay of the parent nucleus is almost an order of magnitude less than this.

This situation is illustrated in the potential energy diagram in Fig 5.20.

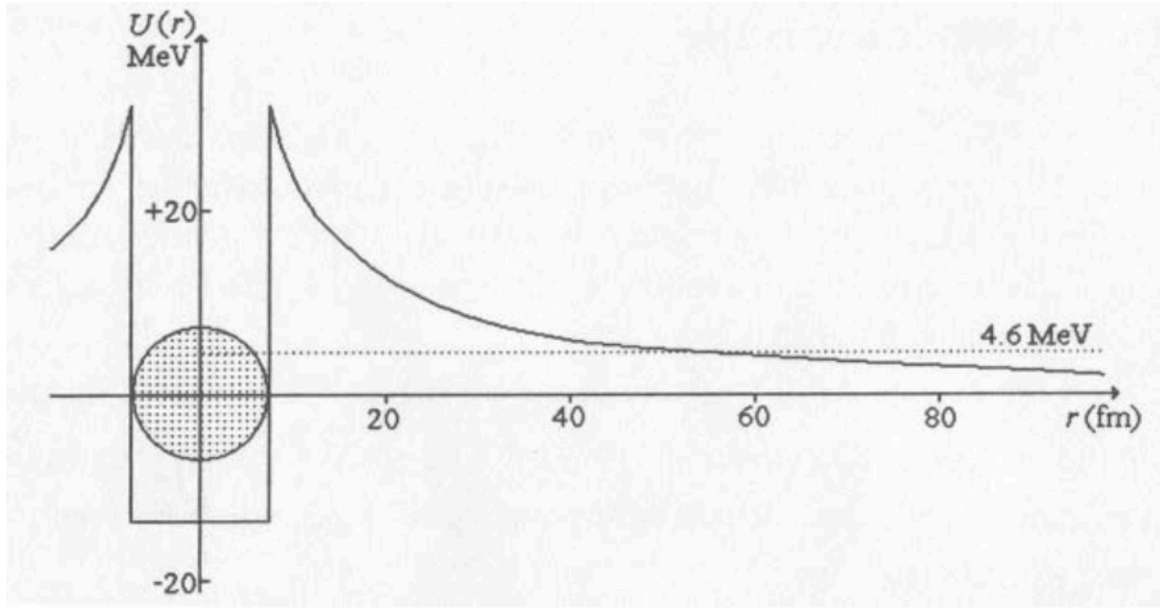


Fig 5.20 The potential energy curve of an a particle and an Rn-222 nucleus. The radius of the nucleus is 7.3fm and the potential energy at its surface is 34MeV. The energy of the a particle, 4.6MeV, is substantially below the maximum height of the potential well enclosing the nucleus.

It would appear that the question to be answered is not why haven't all the α emitting nuclei decayed long ago, but how do any of them manage to emit an α particle at all!

The answer is provided by quantum mechanics.

As we saw in Section 2.4.8, a particle can escape from a finite potential well even though it does not possess sufficient energy to mount its walls; it may simply 'tunnel' through the energy barrier.

The probability of the particle's escape depends principally on three factors: (i) the height of the energy barrier; (ii) its shape; (iii) the energy of the particle.

The emission of an α particle from a nucleus is an example of quantum tunneling.

Thus, the stability of an α emitting nucleus is a function of the escape probability of the α particle.

Generally, this probability is extremely low.

This is why alpha emitting nuclei still occur naturally.

The probability of their decay is so low, that even the eons of time since their creation have not sufficed for all of them to disintegrate.

The stability of a radioactive isotope is reflected by its half-life.

The longer the half-life of an α emitting isotope, the more stable are its nuclei and the lower is the escape probability of the α particles.

The empirical dependence of the half-life of the α emitting isotopes of thorium on the energy, E_α, of the α particle is illustrated in Fig 5.21.

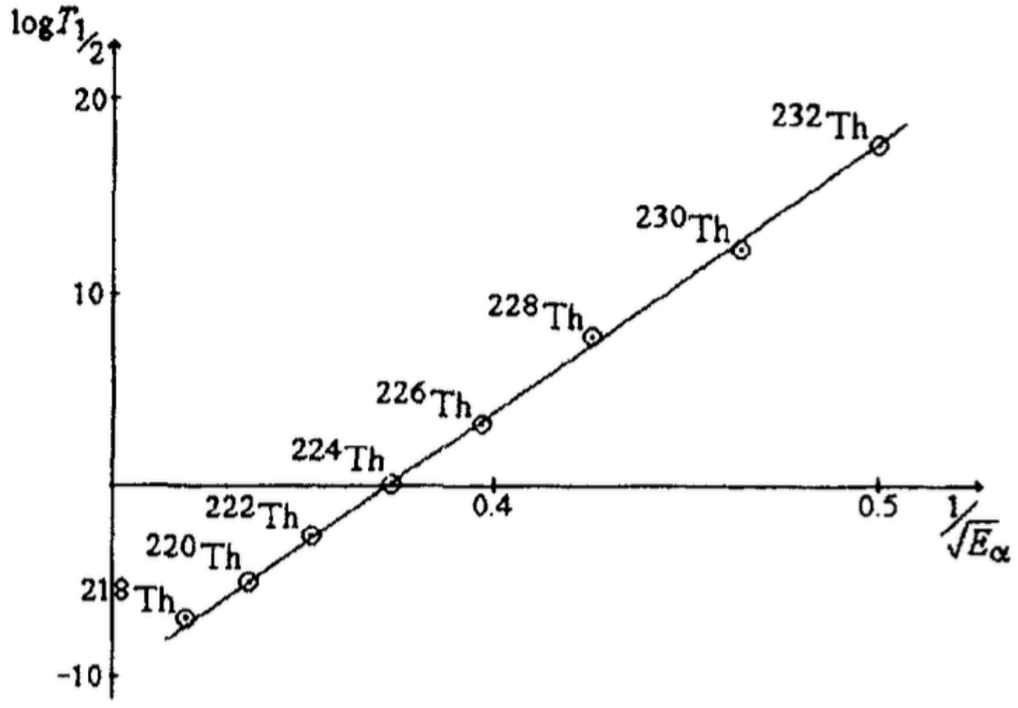


Fig 5.21 The logarithm log $T_{1/2}$, of the half-lives (in seconds) of the α emitting isotopes of thorium as a function of the reciprocal, $1/\sqrt{E_\alpha}$, of the square root of the α particle's energy.

To a close approximation, the escape probability, P_α, of the α particle from a nucleus is an exponential function of the reciprocal of the square root of the α particle's energy:

$$P_\alpha \propto e^{1/\sqrt{E_\alpha}} \tag{5.25}$$

Consequently, small differences in particle energy produce great differences in the escape probability.

This rule was first derived from quantum mechanical considerations by assuming that α particles exist as such inside the nucleus and that they bounce back and forth between the walls of the potential barrier impacting repeatedly on the energy barrier.

Each impact is considered an attempt to escape.

The escape probability is then expressed in terms of the number of times the particle must try before succeeding.

Though, on this view, an α particle typically knocks at its confining wall 10^{21} times every second, it may have to wait millions of years before it escapes.

5.3.7 The Mechanism of β Decay-Weak Charge

Few natural phenomena have generated as many difficulties or have required as many apparently contrived and ad hoc postulates for their elucidation as has the phenomenon of β decay.

On the other hand, few have ultimately proven as fruitful a source of new knowledge and ideas.

An examination of the disintegration energies, Q, of the three basic processes underlying β decay reveals that, without an additional source of energy, neither β+ emission nor electron capture can occur spontaneously.

β^- emission:	$n \rightarrow p + e^-$	$Q_{\beta^-} = +0.78\text{MeV}$
β^+ emission:	$p \rightarrow n + e^+$	$Q_{\beta^+} = -1.80\text{MeV}$
electron capture:	$p + e^- \rightarrow n$	$Q_{EC} = -0.78\text{MeV}$

(5.26)

The endotherrmicity of β+ emission accounts for the stability of free protons and that of electron capture for the existence of hydrogen atoms; the spontaneous fusion of the latter's proton and electron to produce a neutron is energetically forbidden.

It follows that β+ emission and electron capture can only take place if the re-arrangement of the nucleons in the daughter nucleus increases the total binding energy at least by the amount needed to make up for the energy lost in the conversion of the proton into a neutron.

The exothermicity of β- emission indicates that the spontaneous disintegration of free neutrons is energetically possible.

In fact, free neutrons are unstable and disintegrate by β- emission after an average lifetime of about 15 minutes.

The neutrons in a nucleus are only prevented from disintegrating by the absence of available energy levels into which the protons that would be produced from their decay could be placed.

The lower proton energy levels are already occupied by the existing protons in the nucleus and the protons that would be produced lack the necessary energy to occupy the vacant higher levels.

Neither the electrons nor the positrons involved in the three modes of β decay are themselves nuclear components.

Nor, since protons, neutrons, positrons and electrons are all fermions with a spin of 1/2, can any one of them be a combination of any two of the others.

The neutron cannot comprise a proton and an electron as was once thought, since it would then have spin of 1/2 +1/2 = I or 1/2 +(-1/2) = 0.

By the same reasoning, each of the formulations that appear in equation (5.26) for the three basic processes underlying β decay violates angular momentum conservation.

Taking β- emission as the example: whereas the neutron has a spin of 1/2, the sum of the spins of the products of its disintegration, a proton and an electron, must be 1/2 +1/2 = 1 or1/2+(-1/2)= 0.

The problem is compounded by an examination of the energies of the electrons and positrons emitted in β decay (Fig 5.22).

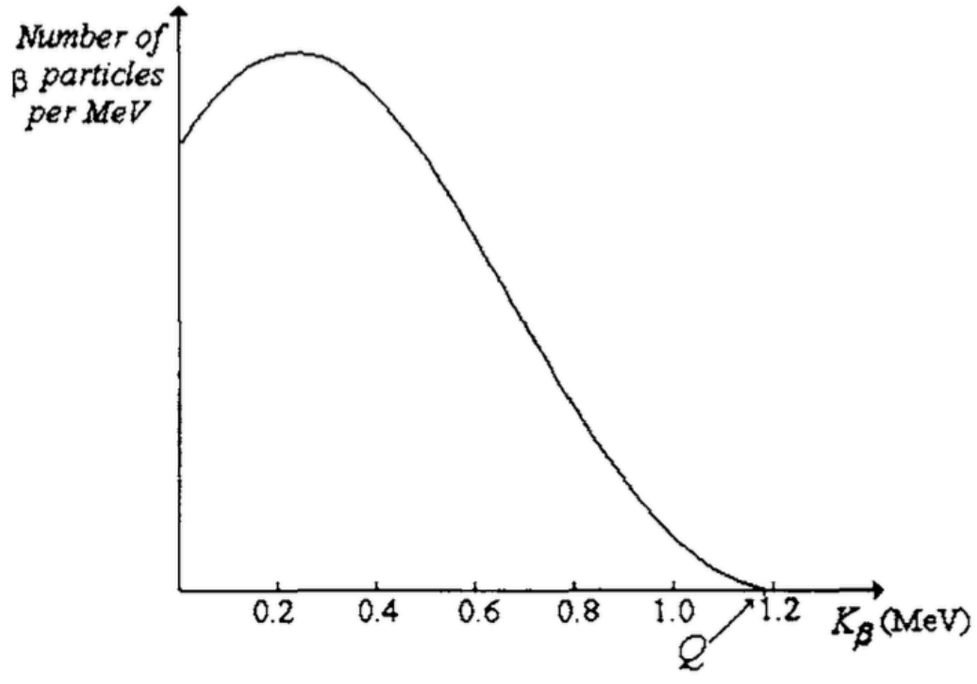


Fig 5.22 Energy spectrum of the electrons emitted from a typical β emitter, Bi-210. Most of the electrons are emitted with energies, K_β, less than half that of the disintegration energy, Q.

Whereas α particles are all emitted with energies equal to or close to the disintegration energy of their decay process, β particles are emitted with a range of energies from almost zero up to a maximum equal to the disintegration energy, Q.

The recoil energy of the nucleus in β decay is negligible, so what happens to the excess energy?

Is energy not conserved in β decay?

In 1931, in an act of faith based on no more than his deep belief in the laws of conservation, Pauli proposed that another undetected particle was involved in β decay.

Fermi later christened the particle the *neutrino*.

However, another twenty years were to pass before experimental evidence for the existence of neutrinos was obtained.

Neutrinos are stable, uncharged and virtually massless ($m_{neutrino} < 0.001\ m_e$) fermions.

They belong to the family of leptons and are the fourth stable fundamental component of ordinary matter.

Both neutrinos, υ, and anti-neutrinos, $\bar{\nu}$, are involved in the three types of β decay.

$$\begin{aligned} &\beta^- \text{ emission:} && n \rightarrow p + e^- + \bar{\nu} \\ &\beta^+ \text{ emission:} && p \rightarrow n + e^+ + \nu \\ &\text{electron capture:} && p + e^- \rightarrow n + \nu \end{aligned} \tag{5.27}$$

The neutrinos and anti-neutrinos provide the spin, 1/2, necessary for angular momentum conservation in β decay.

Moving with a velocity close to that of the speed of light, they likewise account for the energy and linear momentum produced in the decay but not taken up by the other particles.

Because they participate in neither strong nor electromagnetic interactions, neutrinos pass through both space and matter unhindered.

As a result, they generally go undetected. They participate only in processes like those that underlie β decay, processes that involve what is termed the *weak interaction*.

The existence of neutrinos and anti-neutrinos was confirmed experimentally when it was shown that the following two *inverse beta decay* reactions occurred:

$$\begin{aligned} p + \bar{\nu} &\rightarrow n + e^+ \\ n + \nu &\rightarrow p + e^- \end{aligned} \tag{5.28}$$

In order to accomplish the first of these reactions, the intense flux of anti-neutrinos generated by the large-scale beta decay that takes place in a nuclear reactor was aimed into a tank of water.

Protons in the water molecules react with the anti-neutrinos and are converted into neutrons and positrons.

The positrons collide almost at once with electrons in the water.

The positron-electron pairs mutually destruct with the production of a pair of γ-ray photons.

The brief existence of the positrons, and hence the occurrence of the inverse beta decay reaction, was inferred from the detection of these photons.

In 1987 a shower of neutrinos (about twenty altogether) were detected in the radiation reaching the earth from an exploding star -a supernova- observed in the Large Magellanic Cloud, one of the closest galaxies, just 170,000 light-years away.

A particularly intriguing feature of neutrinos and anti-neutrinos is the distinction between them.

Being uncharged particles, the difference between them cannot be one of electric charge.

Instead, it is their helicity that distinguishes between them.

Relative to their direction of motion each spins in a different manner.

Whereas the neutrino spins in a counterclockwise direction -it has negative helicity, the anti-neutrino spins in a clockwise direction -positive helicity.

The significance of this distinction is that the neutrino is the mirror image of the anti-neutrino, and vice versa (Fig 5.23).

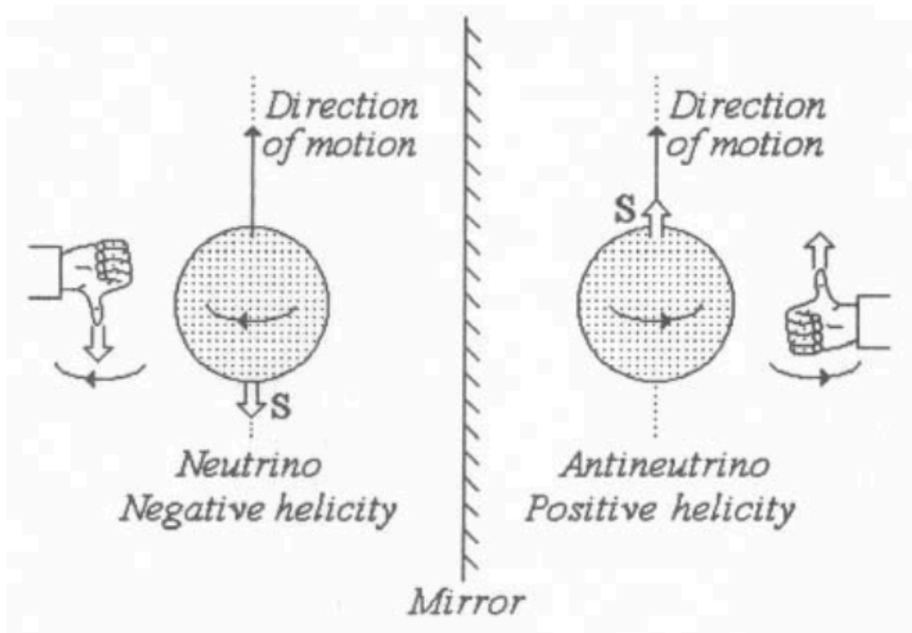


Fig 5.23 The neutrino and anti-neutrino are mirror images of one another. Relative to the axis defined by its direction of motion, the neutrino spins in a counterclockwise direction whereas the anti-neutrino spins in a clockwise direction. The direction of the axial vector, **S**, that represents the particle's spin is given by the right hand (corkscrew) rule.

Helicity is a property of all spin 1/2 particles, such electrons and positrons.

However, unlike the neutrino and antineutrino, electrons and positrons may have either positive or negative helicity, indiscriminately.

According to classical thinking, there is no way of telling whether what we are looking at is an object or its mirror image.

A physical object or process and its mirror image are indistinguishable.

This principle is called *Parity Conservation*. We say that the laws of classical physics are invariant under the parity operation.

Although the world is full of examples of left-right or object-mirror image asymmetries, ranging from the direction of the spirals of certain mollusk shells to the optical activity of many organic molecules, all such examples are thought to be accidental; they do not reflect any inherent asymmetry in the laws of Nature.

The predominance of the L-amino acids in living tissue is considered an evolutionary 'accident', not a requirement for life; in principle a 'mirror image man' constructed of D-amino acids should be a living being.

The parity of the process by which certain unstable nuclei and particles decay was first questioned in 1956 by Lee and Yang.

That parity is conserved in electromagnetic interactions and strong interactions had already been established at the time.

However, it had been found that a certain charged particle, the (positive or negative) kaon, K±, could decay in two different ways, one that produced three pions, the other only two.

In order to account for this anomaly, Lee and Yang postulated that parity is not conserved in processes of this type.

The interaction that underlies β- decay is the same as that involved in the decay of the kaon particle, K±; it is known as the weak interaction.

Lee and Yang associated the non-conservation of parity with this interaction.

On this view, β- decay should also exhibit non-conservation of parity, i.e., the mirror image of β- decay should be different from directly viewed β- decay.

Shortly after Lee and Yang put forward their remarkable hypothesis, this was confirmed in a landmark experiment.

Under suitable circumstances, the spins of certain emitting nuclei such as Co-60, can be aligned by the action of an external magnetic field.

The effect is an indirect one: the field first acts on the atomic electrons and they, in tum, influence the nuclei.

In the mirror image of such a polarized array of nuclei, the orientations of both the magnetic field and the nuclear spin vector are reversed (Fig 5.24).

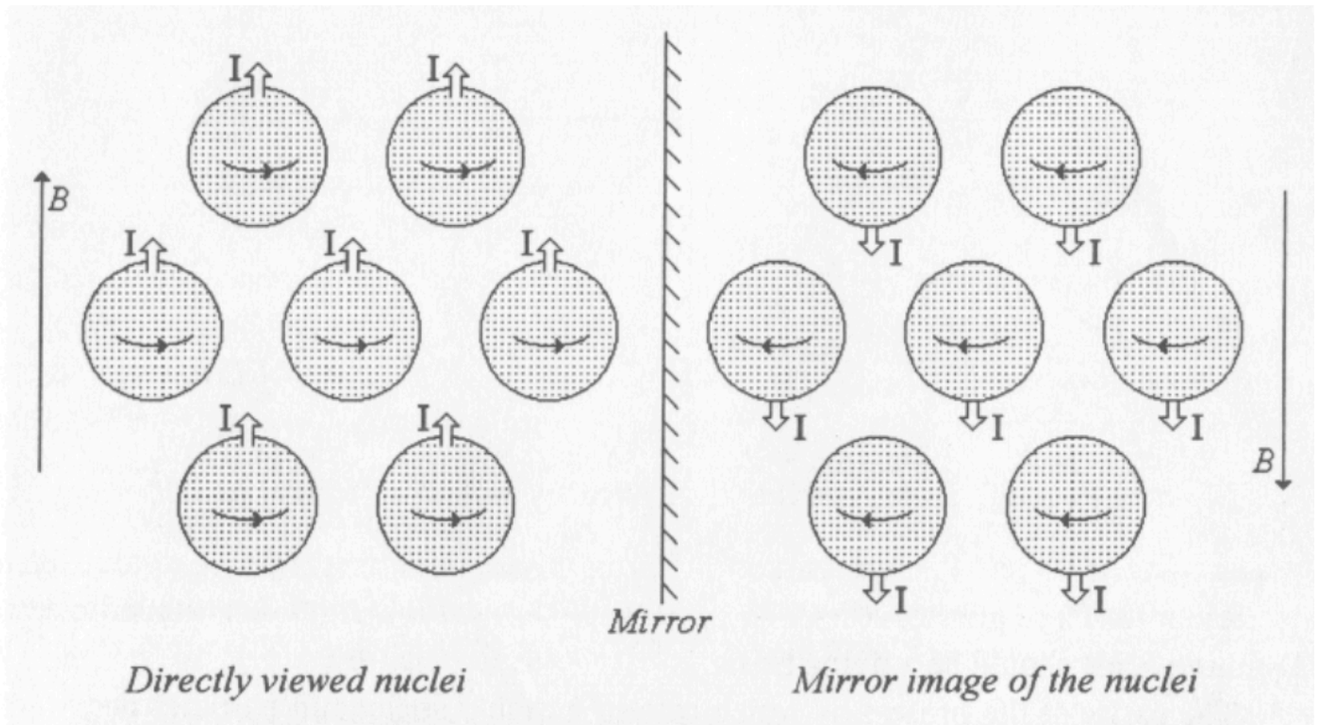


Fig 5.24 A polarized array of Co-60 nuclei. The effect of the magnetic field, B, is to align the nuclear spins of the nuclei. Like all physical quantities represented by axial vectors, in their mirror images, the orientations of the nuclear spins, **I**, and the magnetic field, B, are reversed. Thus, reversing the direction of the external magnetic field and rotating the sample containing the Co-60 nuclei through 180° has the effect of transforming it into its mirror image.

Whatever requirements the conservation of angular momentum imposes on the correlation between the orientation of the spin of the beta electron and that of the emitting nucleus in the polarized array, it can be satisfied whether the electron is emitted in the same direction as the nuclear spin or in the opposite direction.

The only difference is that in the former case the electron's helicity will be positive (right-handed) and in the latter negative (left-handed).

Assuming these options have an equal probability, the same number of electrons should be emitted, on average, in each direction (Fig 5.25).

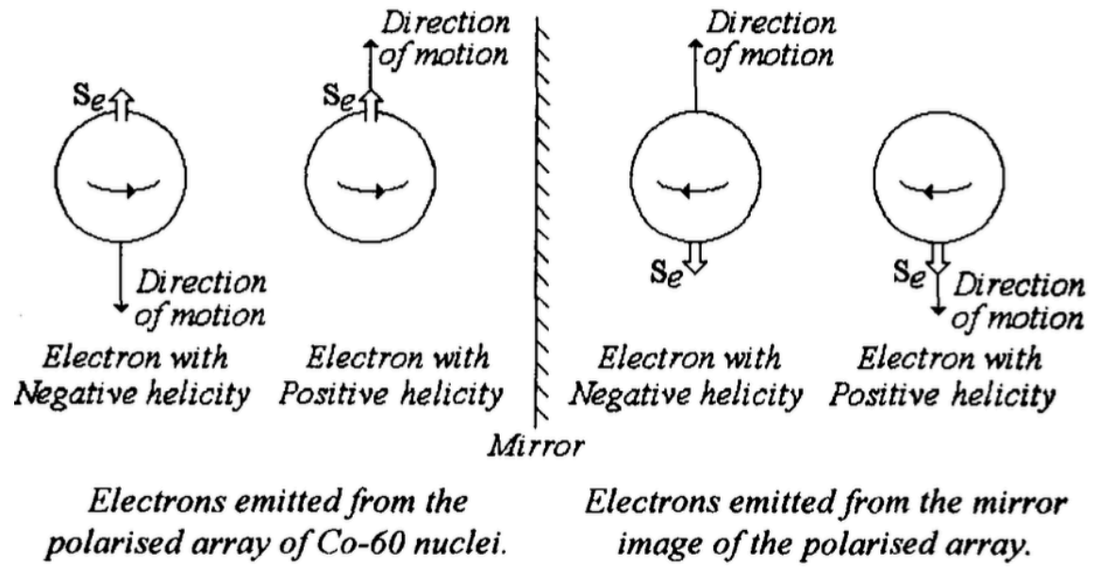


Fig 5.25 The emission of beta electrons from the polarized array of Co-60 nuclei and from its mirror image. The spin of the electrons, $\mathbf{S}_e$, is parallel to that of the nuclear spin. If, on average, the same number of electrons are to be emitted in both directions, the alternative helicities must be equally probable.

The same is true for the mirror image of the array, in which the nuclear spins are pointing in the opposite direction.

Hence, whatever the orientation of the array, we would expect an equal number of beta electrons to be emitted, on average, in each direction.

Experimental confirmation of this would indicate that parity was conserved.

The emission of beta electrons from a polarized array of radioactive nuclei was first studied in an experiment conducted in 1957 at the US National Bureau of Standards by a group led by Chien Shiung Wu.

A sample of the β- emitter Co-60 was placed in a strong magnetic field at a temperature close to absolute zero.

The low temperature served to reduce the random motions of the atoms sufficiently for the magnetic field to align the spin of their nuclei.

It was found, that under these conditions, the sample emitted significantly more beta electrons in the opposite direction to that of the nuclear spins than in the same direction, i.e., more electrons with negative than with positive helicity.

Evidently, decay preferentially produces left-handed electrons. Parity is not conserved in β- decay.

In order to confirm this result, the sample was rotated through 180° and the direction of the magnetic field was reversed.

This has the effect of transforming the apparatus into its mirror image.

Had parity been conserved, the electrons now preferentially emitted would be right-handed ones, i.e., they should be emitted in a direction parallel to that of the nuclear spins (Fig 5.26).

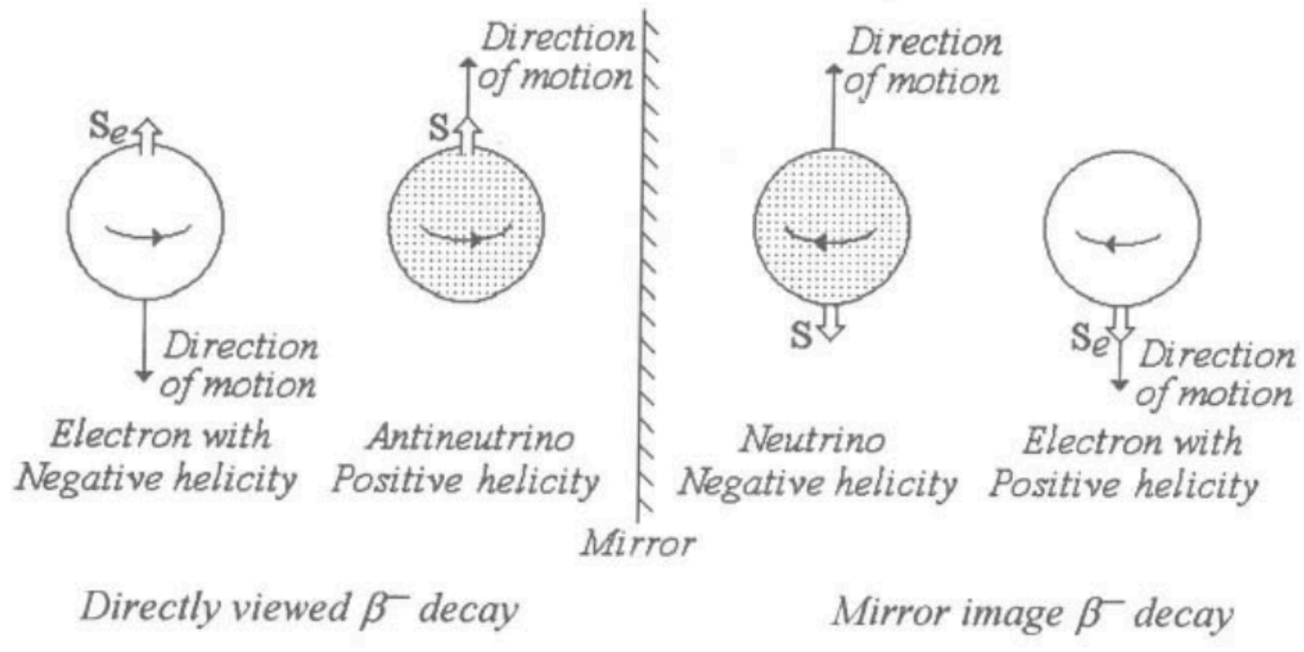


Fig 5.26 β- decay and its mirror image. For parity to be conserved, the electrons emitted from the mirror image sample of Co-60 nuclei should move in the same direction as those emitted from the directly viewed sample and should possess a positive helicity. Likewise, the particles emitted with them should be neutrinos and not antineutrinos. However, the mirror image of β- decay has never been observed.

However, once again the beta electrons were emitted preferentially in the opposite direction to that of the nuclear spins; i.e., the sample produced mainly left-handed electrons.

Evidently, the mirror image β- decay does not occur in nature.

A second experiment was subsequently performed using Co-58, which is a β+ emitter.

In this case the opposite effect was observed, namely, the positrons were preferentially emitted in the same direction as the alignment of the nuclear spins.

This provided additional confirmation of the theory proposed by Lee and Yang and constituted a convincing demonstration of the basic validity of the experiments.

An alternative and perhaps even more intriguing way of relating the neutrino to the anti-neutrino involves time reversal.

This is a particular case of the general rule that because every particle has a probability amplitude for moving backwards in time, every particle has an anti-particle.

When describing the technique of Quantum Electrodynamics, we noted that all the possible paths by which an electron can move between two points in space-time must be taken into account, even those that show the particle moving backwards in time must be included.

Under the right circumstances, the paths of such 'backwards-moving' particles can be long enough to make them observable.

It then transpires that a backwards-moving particle manifests the properties of the ordinary particle's anti-particle.

For example, when viewed by a normal observer, a backwards moving electron attracts ordinary electrons; we say that it carries a positive charge and call it a positron.

On the other hand, there is no difference between a forward and backward moving photon; the photon is its own anti-particle.

On this basis, an anti-neutrino is a neutrino going backwards in time, viewed with time moving forwards.

This strange definition is illustrated in the Feynman diagram in Fig 5.27 which illustrates β- decay in three different but equivalent modes.

In the first all the particles are forward-moving; in the second, a backward-moving positron replaces the electron; in the third, a backward-moving neutrino replaces the anti-neutrino.

The Feynman diagram for β- decay can be transformed into the diagram showing the corresponding inverse beta reactions by the simple expedient of reversing the direction of the world-lines of the anti-neutrino.

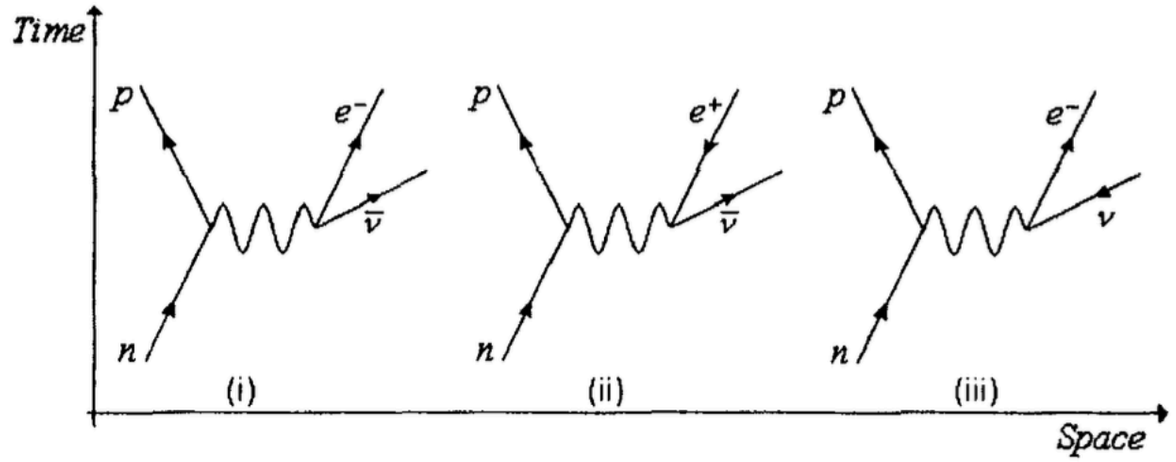


Fig 5.27 Feynman diagrams for the β- decay of a neutron. The three diagrams are different but equivalent: (i) All the particles are forward-moving. (ii) The forward-moving electron is replaced by a backward-moving positron. (iii) The forward-moving anti-neutrino is replaced by a backward-moving neutrino. In each case the decay is mediated by the same photon-like virtual particle.

In terms of the quark hypothesis, each of the three types of β decay involves the transmutation of one first generation quark into another.

In each case, two leptons are also involved, an electron or positron and an anti-neutrino or neutrino, respectively.

$$\begin{aligned} &\beta^- \text{ emission:} && \text{down quark} \rightarrow \text{up quark} + e^- + \bar{\nu} \\ &\beta^+ \text{ emission:} && \text{up quark} \rightarrow \text{down quark} + e^+ + \nu \\ &\text{electron capture:} && \text{up quark} + e^- \rightarrow \text{down quark} + \nu \end{aligned} \tag{5.30}$$

However, since leptons do not participate in the color/strong force, the interaction underlying these transmutations cannot be the color/strong interaction.

It transpires, that in order to account for β decay, a fourth fundamental force, in addition to gravity, electromagnetism and the color (strong) force, must exist in Nature.

This fourth interaction is called the *weak force*; it is weak only relative to the strong force.

It is a very short-range interaction, only operating at distances of up to 10^{-18}m.

The physical origin of the weak force has been designated the weak charge.

This physical property, possessed both by leptons and quarks (A corollary of the parity violation observed in β decay is that the weak charge is only possessed by left-handed particles.

Particles having a right-handed helicity do not participate in the weak interaction), is the source of the weak force just as mass is the source of gravity, electric charge is the source of electromagnetism and color is the source of the color/strong force.

Two kinds of weak charge have been identified.

One, designated u, is the type carried by the up quark, the other, d, is that carried by the down quark.

In terms of the weak charge hypothesis, the three modes of β decay involve the transmutation of one kind of weak charge into another as shown in the following formulations:

$$\begin{aligned} &\beta^- \text{ emission:} && d \rightarrow u + e^- + \bar{\nu} \\ &\beta^+ \text{ emission:} && u \rightarrow d + e^+ + \nu \\ &\text{electron capture:} && u + e^- \rightarrow d + \nu \end{aligned} \tag{5.30}$$

The weak and electric charges of the three generations of leptons and quarks are summarized in Table 5.16.

Table 5.16: The Weak and Electric Charges of the Leptons and Quarks

	Leptons			Quarks		
	Particle	Weak Charge	Charge (e)	Flavour	Weak Charge	Charge (e)
First Generation	Neutrino, ν	u	0	Up, u	u	$+\frac{2}{3}$
	Electron, e	d	-1	Down, d	d	$-\frac{1}{3}$
Second Generation	μ-Neutrino, ν_μ	u	0	Strange, s	u	$-\frac{1}{3}$
	Muon, μ	d	-1	Charm, c	d	$+\frac{2}{3}$
Third Generation	τ-Neutrino, ν_τ	u		Bottom, b	u	$-\frac{1}{3}$
	Tauon, τ	d	-1	Top, t	d	$+\frac{2}{3}$

Like the other fundamental interactions, the weak interaction too can be interpreted in terms of a quantum theory of fields, i.e., in terms of photon-like particles called intermediate bosons, that carry the interaction.

Because weak charge, just like electric charge and color, is conserved in all fundamental interactions, the mediating particles involved in β decay must carry an electric charge.

These are the charged W+ and W- intermediate bosons (Fig 5.28).

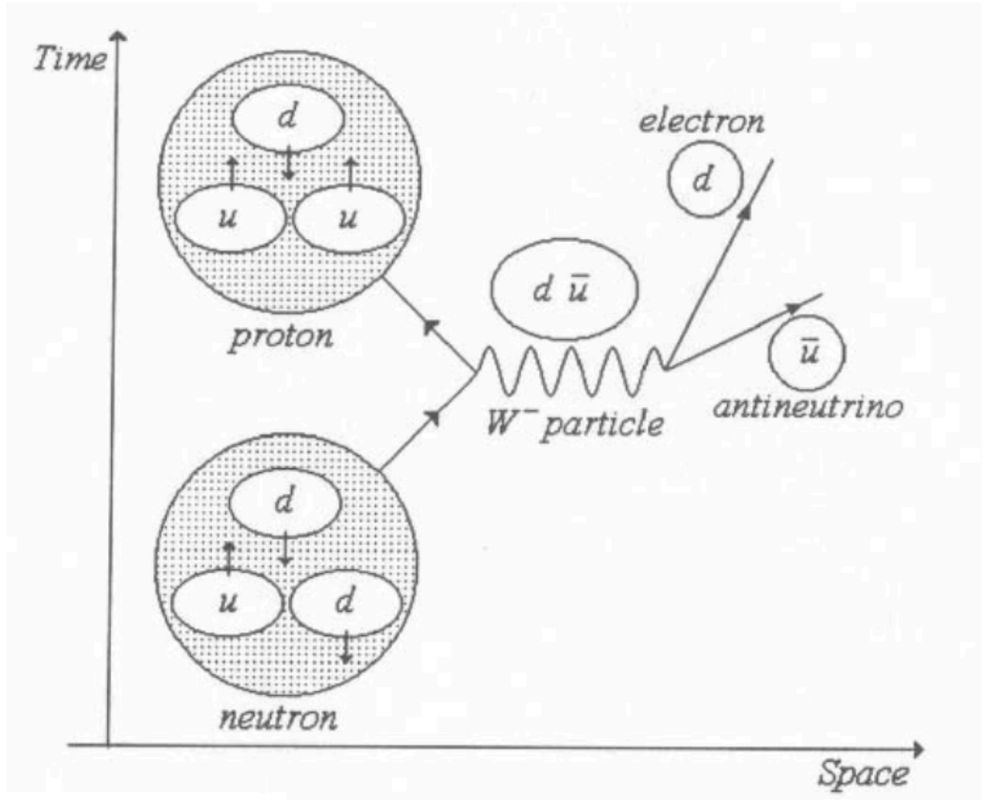


Fig 5.28 The emission of a beta electron viewed as the conversion of a u weak charge into a d weak charge. The process is mediated by the virtual W- particle which comprises a d weak charge and an anti-u weak charge. The conservation of weak charge requires that the electron carry a d weak charge in addition to its negative electric charge and that the anti-neutrino carry an anti u weak charge.

Processes in which like weak charges interact have also been observed; these interactions are mediated by the uncharged Z^0 intermediate boson.

There is an important and fundamental difference between the bosons that mediate the weak and color interactions, the intermediate W+, W- and Z^0 bosons and the gluons, respectively, and the regular photons that mediate electromagnetic interactions.

Whereas the regular photons do not carry any of the electric charge that is the source of the interactions they mediate, the intermediate bosons that mediate the weak force themselves contain weak charge; likewise, the gluons that mediate the color force themselves have color.

Thus, these 'carrier' particles can themselves emit more carrier particles, complicating the situation still further.

Because of the very short-range of the weak interaction, up to distances of just 10^{-18}m, the mediating bosons, W+, W- and Z^0, must all be massive particles.

Assuming, that in weak interactions these particles are exchanged close to the speed of light, their lifetime is of the order of 10^{-26}s.

Thus, their intrinsic energy is approximately:

$$\Delta E = \frac{\hbar}{\Delta t} = \frac{6.58 \cdot 10^{-22}}{10^{-26}} \approx 60{,}000\text{MeV} \tag{5.31}$$

The β decay is a relatively slow process because of the low probability of generating these massive intermediate bosons.

The existence of the W+, W- and Z^0 intermediate bosons was demonstrated in 1984 in experiments between colliding beams of protons and anti-protons carried out at CERN.

They are amongst the heaviest elementary particles ever detected having a mass more than 80 times that of the proton.

Their discovery was a triumph for the Electroweak Hypothesis proposed by Weinberg, Salam and Glashow in 1968.

This hypothesis showed how the weak and electromagnetic interactions could be united into a single electroweak theory in a way similar to that by which Maxwell had integrated electrical and magnetic phenomena into the single framework of electromagnetism, 100 years earlier.

The combination of the electroweak theory with the theory of quantum chromodynamics is referred to in high-energy physics at the Standard Model.

In addition to the leptons and quarks listed in Table 5.16 and the bosons that mediate the interactions between them, the standard model also predicts the existence of a massive particle of intrinsic energy ~1000GeV known as the Higgs boson.

It has now been discovered.

The standard model has proved its usefulness repeatedly and is backed by impressive experimental verification.

The successful unification of the weak and electromagnetic interactions into a single framework has stimulated the search for a theory-a Grand Unifying Theory or GUT-that would unite the four fundamental interactions: gravity and the weak, strong and electromagnetic interactions.

This search, unsuccessful to date, has centered on the symmetries that are thought to underlie all four interactions.

Table 5.17: The Four Fundamental Interactions - The Standard Model

Interaction	Source	Particles Affected	Range	Relative Strength	Mediating Bosons	Role in Nature
Gravity	Mass	All	up to ∞	$\sim 10^{-38}$	Gravitons	Sustains the large-scale structure of the universe
Weak Force	Weak charge; *u* or *d*	Quarks and Leptons	$\sim 10^{-18}$m	$\sim 10^{-13}$	W^+, W^- and Z^0	Stipulates and impels the transformations of quarks and leptons
Electro-magnetism	Electric charge: *e*	Charged Particles	up to ∞	$\sim 10^{-2}$	Photons	Sustains atomic and molecular structures and stipulates the allowed changes in their configurations
Colour Force	Colour: *red*, *blue* or *green*	Quarks	$\sim 10^{-15}$m	1	Gluons	Sustains hadronic and nuclear structures

Part 5.4 Nuclear Reactions and Nuclear Energy

In his New System of Chemical Philosophy, published in 1803, Dalton had proposed that the chemical elements comprise indestructible immutable atoms.

This postulate gave a simple explanation for the failure of the alchemists during the previous thousand years to transmute base metals into gold.

By the base metals and gold were taken to be elements.

According to Dalton, this meant that each comprises its own species of indestructible immutable atoms.

Thus, no amount of heating, cooling, refracting or distilling of this or that mixture of materials could ever convert lead or any other base metal into gold.

Dalton's model implied that any reactions induced by such processes could only change the way the atoms are ordered and grouped within the materials; they could not affect the essential attributes of the atoms themselves.

The discovery of the natural radioactivity of the heavy elements marked the first breach in Dalton's model.

The atoms of these elements are not immutable they undergo spontaneous changes that alter their elemental identity.

The daughter nuclei that result from radioactive decay are invariably those of a different element than the parent nuclei.

The alchemists' dream of inducing the transmutation of one element into another was first realized by Rutherford, fortuitously, in 1919.

It occurred in the course of experiments he was conducting into the 'stopping power' of various gases.

In these investigations, a scintillation counter was used to count the number of α particles that penetrated a given distance through each gas.

Gases such as oxygen and carbon dioxide gave the expected results, viz. the number of scintillations observed at a given distance from the radioactive source diminished when these gases were admitted into the apparatus.

But with the admission of air into the apparatus quite unexpected and singular effects were observed. As Rutherford himself reported:

> *A surprising effect was noticed, however, when dried air was introduced Instead of diminishing, the number of scintillations was increased, and for an absorption corresponding to about 19cm of air the number was about twice that observed when the air was exhausted [removed].*

The additional scintillations were evidently caused by a different particle, not by the α particles, one whose penetrating power was greater.

This suggested that they had a smaller mass.

The 'brightness' of the scintillations produced by these 'long-range' particles was comparable to that produced in the same apparatus by high-energy hydrogen atoms.

Further experiments showed that it was the nitrogen in the air that was responsible for the production of these additional 'long-range' scintillations.

Rutherford wrote:

> *From the results so far obtained it is difficult to avoid the conclusion that the long-range atoms [particles] arising from the collision of α particles with nitrogen are probably atoms of hydrogen ... If this be the case, we must conclude that the nitrogen atom is disintegrated under the intense forces developed in a close collision with a swift α particle, and that the hydrogen atom which is liberated formed a constituent part of the nitrogen atom ... The results as a whole suggest that, if α particles - or similar projectiles - of still greater energy were available for experiment we might expect to break down the nucleus structure of many of the lighter atoms.*

This remarkable insight ushered in the age of nuclear physics.

The reaction between the α particles and nitrogen atoms discovered by Rutherford can be formulated as:

$$\alpha\left({}_{2}^{4}\mathrm{He}^{++}\right) + {}_{7}^{14}\mathrm{N} \rightarrow {}_{8}^{17}\mathrm{O} + p \qquad (5.32)$$

In subsequent experiments, all the other light elements were bombarded with a. particles from natural radioactive sources, and in about ten cases protons were detected.

Protons were clearly a constituent of the nuclei of all these elements.

The ensuing development of particle accelerators, in which the particle projectiles could be accelerated from energies of just a few million electron-volts up to hundreds of million of electron-volts, enabled nuclear reaction experiments to be carried out at much higher energies.

With the use of these particle accelerators, experiments could also be carried out with other particle projectiles, such as protons and gaseous ions.

5.4.1 Nuclear Reactions

For a nuclear reaction to occur, two nuclei, or a nucleus and a particle, must approach to within the range of the strong force they each exert.

Under these circumstances, changes can take place in the way the nucleons are ordered and grouped within the nuclei.

The process is analogous to the changes in the order and grouping of the atoms within a molecule that accompanies a chemical reaction.

The close approach of the two nuclei, or of the nucleus and the particle projectile, is essentially a collision event.

In general, and whether or not a reaction results from their encounter, energy, linear and angular momentum, mass number and electric charge are all conserved in the process.

These close encounters are usually achieved by bombarding target nuclei with particle projectiles.

The process can be formulated as

$$\underset{\textit{Target nucleus}}{X_i} + \underset{\textit{Incident particle}}{x_i} \rightarrow \underset{\textit{Scattered particle}}{x_f} + \underset{\textit{Recoiling nucleus}}{X_f} \qquad (5.33)$$

where the symbols X_i and X_f designate the nuclei and x_i and x_f the particles.

A short-hand designation often used to express nuclear reactions is: $X_i(x_i,x_f)X_f$

If the reaction leaves the identity of the particle projectile unchanged, i.e., $x_i = x_f$, it is called *scattering*.

If kinetic energy is conserved in the reaction the scattering is termed *elastic*.

An example of elastic nuclear scattering is the classic Rutherford experiment in which α particles were scattered unchanged on passing through a gold foil.

If kinetic energy is not conserved, the scattering is termed *inelastic*.

Nuclear reactions generally occur in two stages.

In the first, a compound excited nucleus is formed that comprises all the nucleons and energy of the colliding species.

Typically, this unstable intermediate nucleus is very short-lived, ~10^{-16}s, and so it goes unobserved.

In the second stage, the compound nucleus decays into the final products of the reaction.

For example, in the first nuclear reaction discovered by Rutherford, ^{14}N (α, p)^{17}O, the intermediate compound nucleus is ^{18}F*.

$$\begin{aligned} \alpha \quad + \quad {}^{14}_{7}\mathrm{N} \quad &\rightarrow \quad {}^{18}_{9}\mathrm{F}^{*} \\ {}^{18}_{9}\mathrm{F}^{*} \quad &\rightarrow \quad {}^{17}_{8}\mathrm{O} \quad + \quad p \end{aligned} \tag{5.34}$$

A particular compound nucleus may be produced in more than one way and may also decay by more than one mode.

For example, bombarding ^{63}Cu nuclei with accelerated protons initially produces the compound nucleus ^{64}Zn*.

$$p\left({}^{1}_{1}\mathrm{H}^{+}\right) \quad + \quad {}^{63}_{29}\mathrm{Cu} \quad \rightarrow \quad {}^{64}_{30}\mathrm{Zn}^{*} \tag{5.35}$$

Bombarding ^{60}Ni nuclei with α particles also produces this compound nucleus:

$$\alpha\left({}^{4}_{2}\mathrm{He}^{++}\right) \quad + \quad {}^{60}_{28}\mathrm{Ni} \quad \rightarrow \quad {}^{64}_{30}\mathrm{Zn}^{*} \tag{5.36}$$

This compound nucleus, $^{64}Zn^*$, exhibits three different decay modes.

$$
\begin{aligned}
{}^{64}_{30}\mathrm{Zn}^* &\rightarrow {}^{63}_{30}\mathrm{Zn} + n \\
&\rightarrow {}^{62}_{30}\mathrm{Zn} + n + n \\
&\rightarrow {}^{62}_{29}\mathrm{Cu} + p + n
\end{aligned}
\qquad (5.37)
$$

The probability of each mode is a function of energy but is independent of the route by which the compound nucleus was produced.

Unstable (radioactive) nuclei that do not occur naturally can be manufactured by nuclear reactions.

This artificial radioactivity was first discovered in 1934 by Irene and Frederic Joliot-Curie during their study of the nuclear reactions between α particles and light atoms.

They found that the product of the reaction between the stable isotope of boron, ^{10}B, and α particles was radioactive; it decayed by $\beta+$ emission.

The source of this radiation was found to be a hitherto unknown isotope of nitrogen, ^{13}N.

$$
\begin{aligned}
\alpha\left({}^{4}_{2}\mathrm{He}^{++}\right) + {}^{10}_{5}\mathrm{B} &\rightarrow {}^{14}_{7}\mathrm{N}^* \\
{}^{14}_{7}\mathrm{N}^* &\rightarrow {}^{13}_{7}\mathrm{N} + n
\end{aligned}
\qquad (5.38)
$$

Artificial radioactive isotopes of all the elements have been manufactured.

Many are produced commercially in substantial quantities for use in research and technology.

A particularly interesting group of nuclear reactions are those by which artificial elements possessing an atomic number greater than 92, transuranium elements, are produced.

For example:

$$\begin{aligned} \alpha + {}^{238}_{92}\mathrm{U} &\rightarrow {}^{241}_{94}\mathrm{Pu} + n \\ {}^{241}_{94}\mathrm{Pu} &\rightarrow {}^{241}_{95}\mathrm{Am} + e^- + \bar{\nu} \\ {}^{241}_{95}\mathrm{Am} &\rightarrow {}^{237}_{93}\mathrm{Np} + \alpha \end{aligned} \qquad (5.39)$$

The isotope of plutonium, Pu-241, initially produced by the reaction between the α particles and the uranium nuclei is a β- emitter.

The product of its decay is an isotope, Am-241, of another transuranium element, americium.

This isotope is an α emitter and decays in turn into the relatively long-lived ($T_{1/2} = 2.14 \times 10^6$ years) isotope of neptunium, Np-237.

To date, transuranium elements with atomic numbers up to 119 have been produced.

5.4.2 The Discovery of the Neutron

A year after discovering the nuclear reaction between α particles and nitrogen atoms, Rutherford put forward the hypothesis that nuclei comprise just protons and neutrons.

However, a further 12 years were to pass before the neutron's existence was confirmed experimentally.

Neutrons were particularly hard to observe because they carry no net charge and so are undeflected by either electrical or magnetic fields.

They were also initially confused with high energy photons.

The first step in the discovery of the neutron was made by Bothe in 1930.

He discovered that when the metal beryllium is bombarded with α particles, a radiation of great penetrating power is produced.

This radiation was undeflected by either electrical or magnetic fields and so it was presumed to comprise very high energy photons.

Two years later, Irene and Frederic Joliot-Curie discovered that when the radiation released from the beryllium target was passed through substances rich in hydrogen atoms, like paraffin wax, protons with energies as high as 5.7MeV were released from the material.

Presuming that the protons had been released from the wax by the impact of photons in a kind of proton Compton effect, the Joliot-Curies calculated that the energy of the photons in the mysterious radiation was about 50MeV.

However, "it is difficult to account for the production of a quantum of 50MeV from the interaction of a beryllium nucleus with an α particle of kinetic energy of 5MeV".

The mystery was solved by James Chadwick in 1932.

Applying elementary mechanics to the problem he suggested that:

> *If we suppose that the radiation is not a quantum radiation, but consists of particles of mass very nearly equal to that of the proton, all the difficulties connected with the collisions disappear.*

A corollary of the classic theory of elastic collisions is that when a stationary particle is struck by a moving particle of the same mass, the moving particle stops dead and the stationary one moves off at the velocity with which it was struck.

In the collision, all the kinetic energy of the incident particle is transferred to the target particle.

Accordingly, if the particles striking the protons in the paraffin wax have nearly the same mass as the protons, their kinetic energy need be no more than 5.7MeV (and not some ten times as much as the Joliot-Curies had supposed) in order that a proton acquire a kinetic energy of 5.7MeV as a result of a collision.

Chadwick continued:

> *In order to explain the great penetrating power of the radiation we must further assume that the particle has no net charge. We may suppose it (to be) the 'neutron' discussed by Rutherford in his Bakerian Lecture of 1920. It is concluded that the radiation consists, not of quanta as hitherto supposed, but of neutrons, particles of mass 1, and charge 0.*

Thus, the reaction between the α particles and the beryllium can be formulated as:

$$\alpha \left({}_{2}^{4}\mathrm{He}^{++} \right) + {}_{4}^{9}\mathrm{Be} \rightarrow {}_{6}^{13}\mathrm{C}^{*}$$
$$ {}_{6}^{13}\mathrm{C}^{*} \rightarrow {}_{6}^{12}\mathrm{C} + n \qquad (5.40)$$

Other methods of producing neutrons were soon discovered.

For instance, it was found that the α particles could be replaced by protons or positive ions that had been accelerated in cyclotrons or by the high potentials of a van de Graff generator.

Moreover, light elements other than beryllium could be used as the target for these projectiles.

By means of these alternative techniques, neutrons in a whole range of different energies have been produced.

Having no electrostatic repulsion to overcome, neutrons can enter the range of the strong force exerted by a nucleus with ease.

In a close encounter with a nucleus, a neutron may be scattered, elastically or inelastically, or it may be captured by the nucleus.

The capture of a neutron does not alter the chemical identity of the nucleus; it just converts it into a heavier isotope of the same element.

For example, irradiating stable Al-27 nuclei with neutrons produces unstable (radioactive) Al-28 nuclei.

$$n + {}^{27}_{13}\mathrm{Al} \rightarrow {}^{28}_{13}\mathrm{Al} + \gamma \qquad (5.41)$$

In a similar way, almost every stable isotope can be converted into an artificial radioactive isotope.

Such reactions are called radiative capture.

They are the basis of the commercial production of the short-lived artificial isotopes such as I-131 and Tc-99 used for medical diagnoses. Nuclear (atomic) reactors, that release an abundance of neutrons, are the neutron source usually employed.

Neutrons do not themselves ionize the atoms and molecules with whose nuclei they collide.

However, the products of their reactions with the nuclei may include ionizing radiations.

For example, a particles are produced in the reaction $^{10}B(n, \alpha)^{7}Li$ and high energy photons (γ rays) are released in radiative capture reactions.

The various devices used to detect neutrons operate by registering the appearance of these ionizing radiations.

5.4.3 Nuclear Cross-sections

In practice, inducing nuclear reactions is very much a hit or miss affair.

In the hope that some of the nuclei and particle projectiles will collide and react, a target containing a multitude of the nuclei is placed in the path of a beam of the particles or is irradiated by a shower of them.

Very few reactions are actually observed.

For instance, in 1925 an attempt was made to photograph the tracks left in a cloud chamber by the particles involved in the first nuclear reaction discovered by Rutherford.

Some 20,000 photographs were taken.

On average, each photograph revealed the tracks of 20 α particles making a total of 400,000 α particle tracks overall.

However, clear signs of a reaction between an α particle and a nitrogen atom in the atmosphere inside the cloud chamber were detected in just eight photographs.

The probability that a nuclear reaction will occur is expressed in terms of a parameter called the nuclear cross-section; the larger the nuclear cross-section the more probable the reaction.

This factor can be viewed as the target area effectively presented to the incident particles by each nucleus (Fig 5.29).

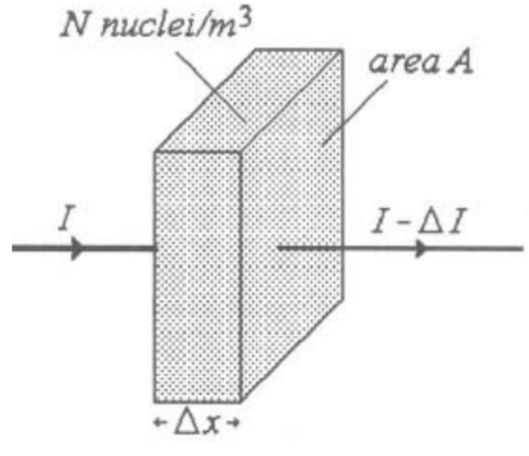


Fig 5.29 The attenuation of a beam of particles by a slice of a material containing N reacting nuclei per cubic meter. The slice effectively presents an impervious 'wall' of area $AN\sigma\Delta x$ to the beam.

Nuclear cross-sections are measured in a unit called the *barn*: lb= $10^{-28}m^2$.

They may be larger or smaller than the geometric cross-section of the target nucleus, though in most cases they are of the same order of magnitude.

In general, the reduction, ΔI, in the intensity of a particle beam of intensity I on passing through a slice of homogeneous material of thickness, Δx is given by

$$\Delta I = -\mu I \cdot \Delta x \tag{5.42}$$

μ is the *macroscopic cross-section* or *linear attenuation coefficient*.

If each unit volume of the material contains N nuclei that can interact with the particles in the beam and if each such nucleus presents a target of area a to the incident particles, then a slice of frontal area A effectively presents a 'wall' of interacting nuclei of area $AN\sigma\Delta x$ to the beam (Fig 5.29).

The relative attenuation, $\Delta I/I$ of the beam resulting from its passage through the slice will be:

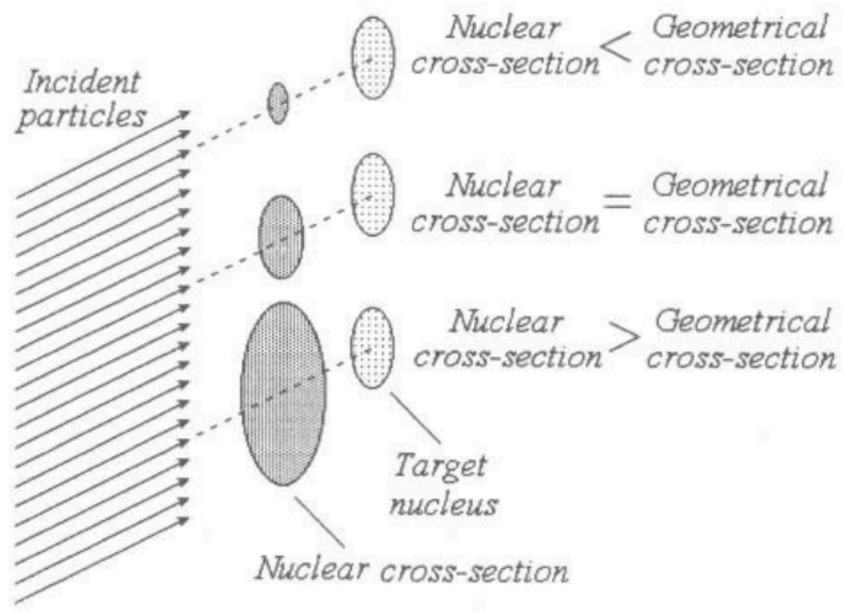
$$\Delta I/I = -AN\sigma\Delta x/A = -N\sigma\Delta x \tag{5.43}$$

From equations (5.42) and (5.43), the relationship between the nuclear cross-section, σ, and the macroscopic cross-section, μ, is simply:

$$\mu = N\sigma \tag{5.44}$$

On average, a particle entering the slice travels a distance of $1/N\sigma$ before reacting with a nucleus.

Fig 5.30 Nuclear cross-sections viewed as the area effectively presented to the incident beam by each target nucleus. The larger the nuclear cross-section the greater the probability that an incident particle will react with a nucleus.



In general, nuclear cross-sections generally depend on two factors:

(i) the energy of the particle projectile;

(ii) the nature of the nuclear reaction.

The dependence of cross-sections on the energy of the particle projectile is striking.

Whereas, in reactions involving positively charged particle projectiles such as protons and a particles, nuclear cross-sections generally increase as the projectiles' energy rises, the opposite is true in the case of neutrons.

This key discovery was made almost inadvertently by Fermi in 1934.

With "no conscious prior reasoning", he decided one day to place a piece of paraffin wax between his neutron source and the target he was irradiating.

The effect was dramatic.

The activity of the irradiated sample increased dramatically, indicating that the neutrons 'filtered' through the wax had induced many more radiative capture reactions of the type shown in equation (5.41) than had the unfiltered neutrons.

Returning from lunch, Fermi explained the phenomenon: the high energy neutrons emitted by the source were being slowed down as they passed through the paraffin wax.

As Chadwick had previously pointed out, when high energy neutrons collide with the hydrogen nuclei in the paraffin wax they are slowed almost to a standstill.

After repeated collisions, they enter into thermal equilibrium, settling down to a kinetic energy of about 0.025eV, this being the average energy of a gas particle at room temperature.

These slow neutrons, called *thermal neutrons*, remain in the vicinity of the target nucleus for much longer than fast neutrons and so the probability of their capture is correspondingly greater.

An alternative equivalent explanation· is that the slower the neutron, the smaller is its linear momentum and the longer its De Broglie wavelength.

However, a particle's De Broglie wavelength is a measure of its size, and so slow neutrons are effectively larger than fast ones and are thus more easily captured by the target nuclei.

Much can be learnt about nuclear structure and nuclear energy levels from the dependence of nuclear cross-sections on the energy of the neutrons.

At certain *resonance energies*, relatively large nuclear cross-sections are observed indicating a high probability that neutrons with these energies will react with the nucleus.

These resonance energies usually correspond to the energy required to raise the target nucleus to one of its excited states.

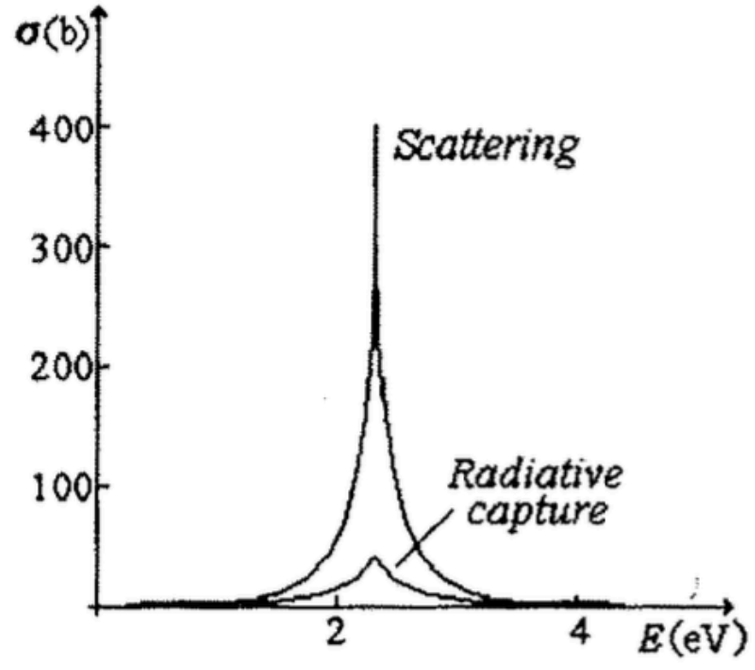
Thus, many nuclear scattering reactions exhibit resonances analogous to those displayed by the mercury atoms in the Franck-Hertz experiment.

For example, both the radiative capture, $^{123}Te(n;\gamma)^{124}Te$, and scattering, $^{123}Te(n,n)^{123}Te$, reactions of neutrons with the tellurium isotope Te-123 exhibit striking resonances with neutrons of energy 2.33eV.

The nuclear cross-section for the former is 40b and that for the latter is 400b.

This is attributed to the presence of a 2.33eV excited energy state in the tellurium nucleus (Fig 5.31).

Fig 5.31 The observed nuclear cross sections, σ, as a function of the neutron energy, E, for the radiative capture, $^{123}Te(n,\gamma)^{124}Te$, and scattering, $^{123}Te(n,n)^{123}Te$, reactions. The latter reaction proceeds via the formation of a compound nucleus, $^{124}Te^*$, which subsequently decays with the emission of a neutron.



5.4.4 Nuclear Energy - Fusion and Fission

The energy changes involved in nuclear reactions such as

$$\underset{\substack{\textit{Target}\\\textit{nucleus}}}{X_i} + \underset{\substack{\textit{Incident}\\\textit{particle}}}{x_i} \rightarrow \underset{\substack{\textit{Scattered}\\\textit{particle}}}{x_f} + \underset{\substack{\textit{Recoiling}\\\textit{nucleus}}}{X_f}$$

can generally be studied by the familiar tools of classical mechanics.

Applying, respectively, linear momentum and energy conservation to this reaction gives:

$$\sum \vec{p}_i = \sum p_f \tag{5.46}$$

and

$$Q + \sum K_i = \sum K_f \tag{5.47}$$

where Q is the difference between the final, $\sum K_f$, and initial, $\sum K_i$, kinetic energies of the system.

The value of Q is independent of the reference frame from which the linear momenta and kinetic energies are measured and is given by the formula:

$$Q = \left[\left(M_i + m_i\right) - \left(M_f + m_f\right)\right]c^2 \tag{5.48}$$

where M_i and M_f represent, respectively, the masses of the target and recoiling nuclei and m_i and m_f the masses of the incident and scattered particles.

In an elastic collision Q = O; in this case no change occurs in the internal energy of the system.

When Q > 0, the reaction releases kinetic energy at the expense of the system's internal energy; the reaction is *exoergic*.

When Q < 0, the reaction requires an investment of energy in order to raise the system's internal energy; the reaction is *endoergic*.

The value of Q can also be calculated from the average binding energy per nucleon, E_B/A of the species reacting and produced in the reaction (Fig 5.32).

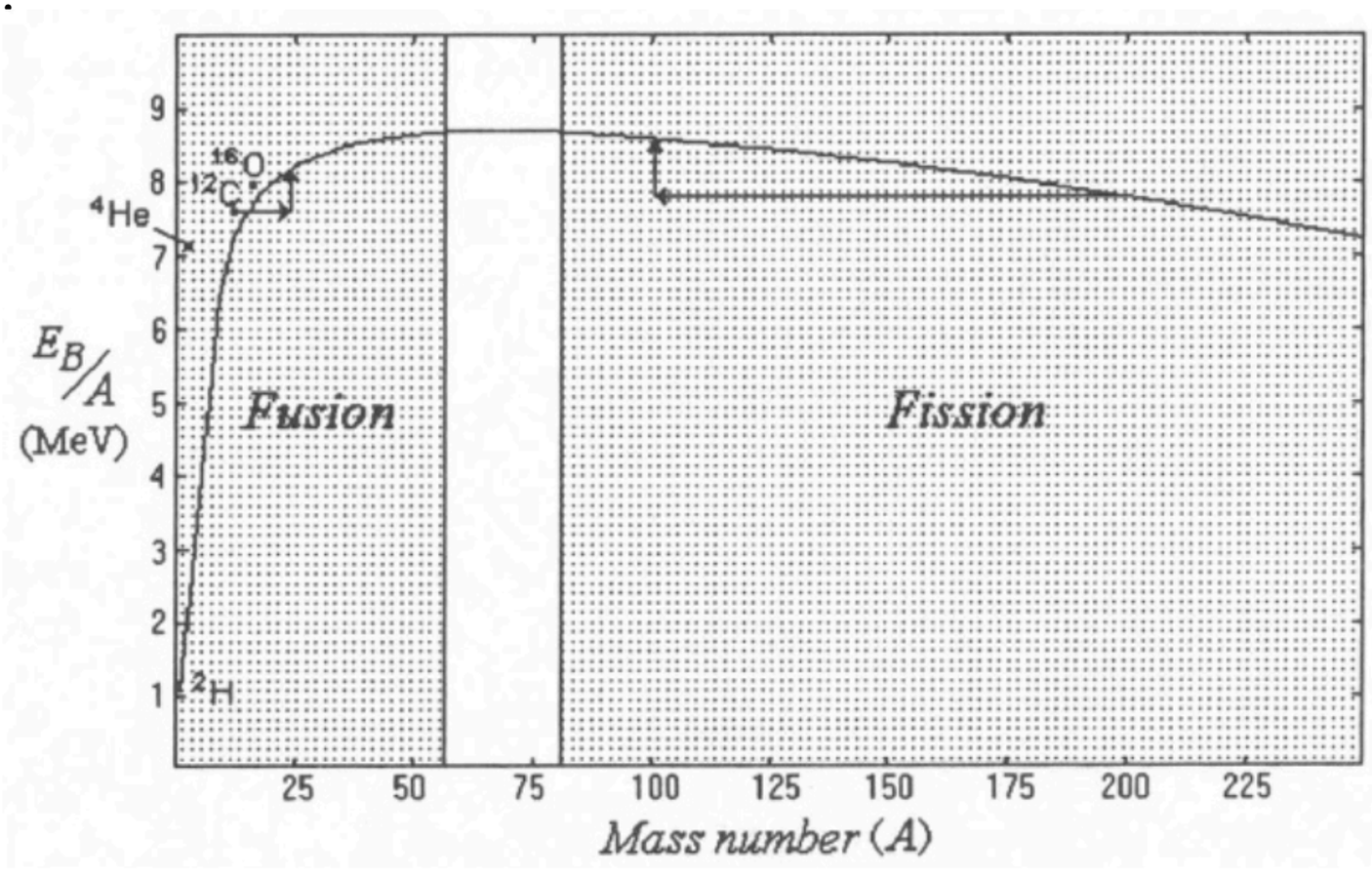


Fig 5.32 The average binding energy per nucleon, E_B/A, as a function of the mass number, A. New nuclei in which the binding energy per nucleon is greater than it was in the original nuclei are produced by both the fusion of two light nuclei (A< 50) and the fission of a heavy nucleus (A> 80).

In general, if the average binding energy per nucleon is higher in the products of a nuclear reaction than in its reactants, Q will be positive and the reaction will be *exoergic*.

If it is lower, Q will be negative, and the reaction will be *endoergic*.

An examination of the graph in Fig 5.32 suggests two possible nuclear processes in which the products would have a greater average binding energy per nucleon than the reactants.

One is the fusion of light nuclei, the other is the fission of heavy nuclei.

For instance, the fusion of two ${}^{12}_{6}C$ nuclei into a single ${}^{24}_{12}Mg$ nucleus would increase the average binding energy per nucleon from 7.7MeV to 8.3MeV, i.e., by 0.6MeV per nucleon.

Since a total of 24 nucleons would be involved in this fusion reaction, the fusion would lead to the release of about 15MeV of energy.

The average binding energy per nucleon would also be increased if a heavy nucleus was to split into two medium-sized nuclei.

For instance, the symmetrical fission of a nucleus containing 200 nucleons into two medium-sized nuclei each of mass number 100 would increase the average binding energy per nucleon from 7.9MeV to 8.7MeV, leading to the release of some 160MeV of energy.

Furthermore, whereas the ratio, N/Z, between the number of its neutrons and protons in nuclei containing 200 nucleons is about 1.5, its value in stable medium-sized nuclei is less than 1.4.

Thus, in addition to the energy released, the fission of heavy nuclei should be accompanied by a shower of free neutrons.

The demarcation line between those nuclei that can in theory fuse exoergically and those that can fission exoergically is marked by the band of medium-sized nuclei whose mass numbers are between 55 and 85, in which the average binding energy per nucleon rises to its maximum value, ~8.8 MeV.

5.4.5 Nuclear Fusion

The fusion of light nuclei is the primary energy producing process in the observed universe; it is the power source that lights the stars and provides the sun with its energy.

On earth, the uncontrolled (explosive) fusion of light nuclei was first achieved in 1952 with the detonation of the first hydrogen bomb. Controlled non-destructive fusion has yet to be accomplished.

In order for fusion to occur, the nuclei must first approach to within the range of the strong force each exerts, viz. to an internuclear distance less than 10^{-14}m.

They must then remain within this range long enough for the strong interaction between them to fuse them into a new and larger nucleus.

Bringing a nucleus to within 10^{-14}m of another nucleus, requires the investment of a considerable amount of energy in order to overcome the long-range electrostatic repulsive forces that act between them.

For example, at a separation of 10^{-14}m, the electrostatic potential energy, U_E, of a pair of protons is:

$$U_E = \frac{ke^2}{r} = \frac{9 \cdot 10^9 \left(1.6 \cdot 10^{-19}\right)^2}{10^{-14}\left(1.6 \cdot 10^{-13}\right)} = 0.144\text{MeV} \tag{5.49}$$

One possible way of supplying this energy is to raise the temperature of the nuclei.

Assuming that the classical equation, $\bar{E} = kT$, for the average energy of an oscillator remains valid at very high temperatures, the average kinetic energy of a gas particle reaches 0.144MeV at a temperature of ~10^9K.

Assuming that the kinetic energy has a Maxwellian distribution, some particles will possess this energy momentarily at a temperature of ~l0^8K.

Quantum mechanical considerations suggest that even at a temperature of ~10^7K particles could 'tunnel' through the electrostatic energy barrier and fuse (Fig 5.33).

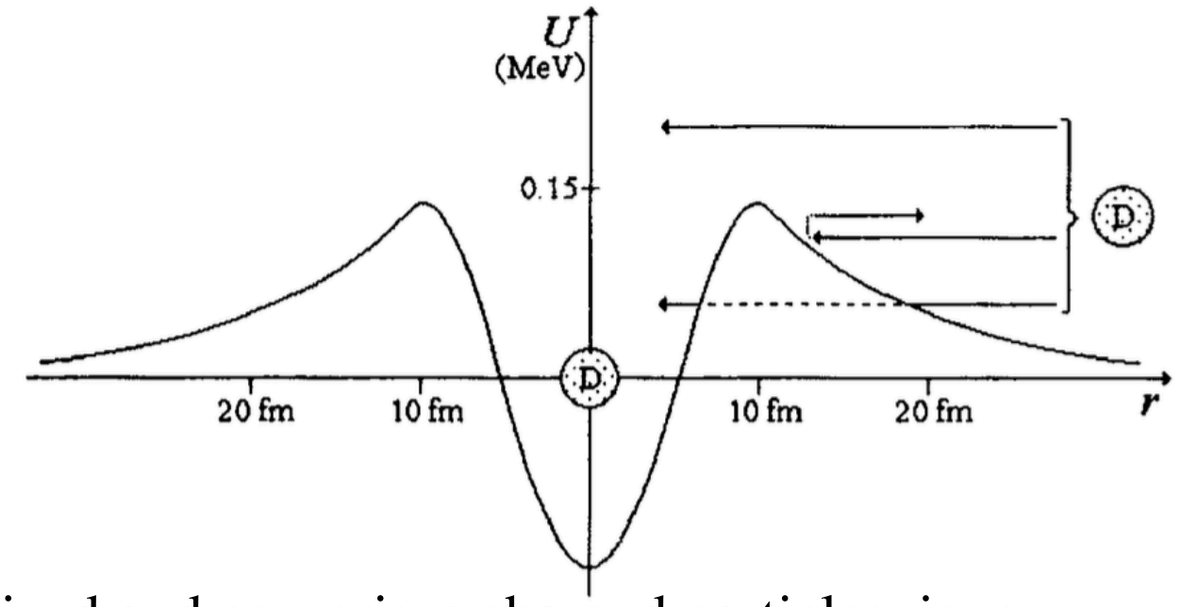


Fig 5.33 Potential energy diagram of two deuterons. In order to clear the potential barrier and fuse with the target deuteron, an oncoming deuteron needs an energy greater than 0.144MeV Most deuterons possessing less energy will be reflected from the barrier though some may tunnel through.

At these extreme temperatures, the gas is completely ionized and comprises charged particles, ions (nuclei) and electrons; in this state it is called a plasma.

Attaining these temperatures is particularly difficult because of the fundamental principle, expressed in the Stefan-Boltzmann Law, $E \propto T^4$, that the energy density of a system varies as the fourth power of its absolute temperature.

Hence, doubling the temperature requires sixteen times the energy density.

Uncontrolled (explosive) fusion is the process by which the visible stars produce energy; it is also the process employed in thermonuclear weapons.

In thermonuclear weapons, the extreme initial conditions required for the fusion of the deuterium and tritium nuclei that comprise their 'fuel' are achieved by using a fission bomb to initiate the reaction.

At the great densities and pressures found at the core of stars such as the sun, hydrogen is 'burnt' at a temperature of about 10^7K in nuclear reactions that convert it into helium.

$$
\begin{aligned}
{}^{1}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H} &\rightarrow {}^{2}_{1}\mathrm{H} + e^{+} + \nu \qquad Q = 1.19\mathrm{MeV} \\
{}^{1}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} &\rightarrow {}^{3}_{2}\mathrm{He} + \gamma \qquad Q = 5.49\mathrm{MeV} \\
{}^{3}_{2}\mathrm{He} + {}^{3}_{2}\mathrm{He} &\rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H} \qquad Q = 12.85\mathrm{MeV}
\end{aligned}
\tag{5.50}
$$

The process is called the *proton-proton cycle*.

The net effect of these reactions is the conversion of four hydrogen nuclei (protons) into a helium nucleus (an a. particle), two positrons and two neutrinos.

The positrons collide almost at once with electrons in the plasma, mutually annihilating and releasing two γ photons.

The neutrinos fly away unhindered, radiating away from the sun together with the solar energy; a flux of about 10^{15} solar neutrinos per square meter per second has been detected on earth.

Note, that no neutrons are produced in the process; such fusion reactions are termed *aneutronic*. Life on earth would be impossible if the sun emitted a flux of high energy neutrons.

An alternative aneutronic process by which the fusion of four hydrogen nuclei into a helium nucleus can occur in stars was proposed by Bethe.

In this process, called the carbon cycle, the C-12 nucleus fulfills the role of a catalyst, being consumed in the first stage and regenerated in the final stage of the series of reactions.

$$
\begin{array}{lllllllll}
{}^{1}_{1}\mathrm{H} & + & {}^{12}_{6}\mathrm{C} & \rightarrow & {}^{13}_{7}\mathrm{N} & + & \gamma & & Q = 1.95\mathrm{MeV} \\
 & & {}^{13}_{7}\mathrm{N} & \rightarrow & {}^{13}_{6}\mathrm{C} & + & e^{+} & + \ \nu & Q = 2.22\mathrm{MeV} \\
{}^{1}_{1}\mathrm{H} & + & {}^{13}_{6}\mathrm{C} & \rightarrow & {}^{14}_{7}\mathrm{N} & + & \gamma & & Q = 7.54\mathrm{MeV} \\
{}^{1}_{1}\mathrm{H} & + & {}^{14}_{7}\mathrm{N} & \rightarrow & {}^{15}_{8}\mathrm{O} & + & \gamma & & Q = 7.35\mathrm{MeV} \\
 & & {}^{15}_{8}\mathrm{O} & \rightarrow & {}^{15}_{7}\mathrm{N} & + & e^{+} & + \ \nu & Q = 2.71\mathrm{MeV} \\
{}^{1}_{1}\mathrm{H} & + & {}^{15}_{7}\mathrm{N} & \rightarrow & {}^{12}_{6}\mathrm{C} & + & {}^{4}_{2}\mathrm{He} & & Q = 4.96\mathrm{MeV}
\end{array}
\qquad (5.51)
$$

When all its hydrogen has been converted into helium, the star contracts, heating up until it reaches the temperature at which three helium nuclei can fuse into a carbon nucleus.

After all its helium has been exhausted, the process is repeated, the star contracting, the temperature rising and heavier nuclei being synthesized until the Fe-56 nucleus is reached.

At this stage fusion ceases to be exoergic.

It is hypothesized, that the heavier elements, $A > 56$ and $Z > 26$, are then synthesized from the medium-sized nuclei by a series of repeated neutron captures and β- decays, the former increasing the mass number of the nuclei and the latter their atomic number.

The requisite neutrons are initially produced in the mature stars by reactions such as

$$\begin{aligned} {}^{4}_{2}\mathrm{He} + {}^{13}_{6}\mathrm{C} &\rightarrow {}^{16}_{8}\mathrm{O} + n \\ {}^{4}_{2}\mathrm{He} + {}^{17}_{8}\mathrm{O} &\rightarrow {}^{20}_{10}\mathrm{Ne} + n \end{aligned} \tag{5.52}$$

and ultimately, after it has exhausted all its fusion reserves, by the star's final cataclysmic implosion (nova or supernova).

The remaining debris blasts off into space to become *inter alia* the raw material for the planets, moons and asteroids.

Reactions considered possible for the controlled fusion of light nuclei include:

$$\begin{aligned} {}^{2}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} &\rightarrow {}^{3}_{2}\mathrm{He} + n \qquad Q = 3.3\mathrm{MeV} \\ {}^{2}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} &\rightarrow {}^{3}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H} \qquad Q = 4\mathrm{MeV} \\ {}^{3}_{1}\mathrm{H} + {}^{2}_{1}\mathrm{H} &\rightarrow {}^{4}_{2}\mathrm{He} + n \qquad Q = 17.6\mathrm{MeV} \end{aligned} \tag{5.53}$$

The basic fuel for these reactions is the deuterium nucleus, $^{2}\mathrm{H}$, the isotope that constitutes about 0.015% of the naturally occurring hydrogen on the earth's surface.

The abundance of this isotope in the water of the oceans represents a virtually inexhaustible supply of fuel.

Thus, induced controlled fusion reactions could supply all of man's future energy needs.

Confining the plasma at a sufficiently high density long enough for the fusion reaction to occur is the major unsolved problem in the search for controlled fusion.

Obviously, at such extreme temperatures, material containers would be useless.

One solution being tried is to confine the plasma along the axis of a toroidal (doughnut shaped) vessel by means of magnetic fields.

However, sustaining the stability of such 'magnetic bottles' has proved very difficult.

A second technique being tried is the sudden compression of solid fusible material, deuterium-tritium pellets.

It is hoped that by 'zapping' the pellets simultaneously from all sides with the high energy particles in intense laser or particle beams, the material's density will be raised abruptly by a factor of a thousand and its temperature to over ten million degrees.

Under these conditions it is anticipated that fusion will occur. However, at the date of writing, neither technique has produced sustained controlled fusion, despite the huge investment made into their research.

A further problem is that the deuterium-deuterium and deuterium-tritium fusion reactions that have been investigated so far produce a flux of high energy neutrons.

Regulating this flux is virtually impossible because it is unaffected by electric or magnetic fields.

Extracting the energy of these neutrons in order to use it for generating electricity, without producing substantial quantities of unwanted and contaminating radioactive isotopes, also appears to be a major problem.

In the light of this, it has been suggested that though they require a higher temperature for their initiation, other aneutronic reactions should be investigated, such as

$$^{1}_{1}\mathrm{H} + {}^{11}_{5}\mathrm{B} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{4}_{2}\mathrm{He} + {}^{4}_{2}\mathrm{He} \qquad (5.54)$$

5.4.6 Nuclear Fission

The nuclear fission of heavy nuclei was first observed by Hahn and Strassman in 1938, when, they detected barium ions in a solution of uranium salts they were irradiating with slow neutrons,.

The appearance of the barium, whose atomic number 56 is about half that of uranium, was a complete mystery.

Its presence in the irradiated solution was inexplicable in terms of the known modes of radioactive decay.

The riddle was solved just a few weeks after the discovery by Meitner and Frisch. Their explanation was based on the liquid-drop model of the nucleus.

> *They pictured the uranium nucleus as a liquid drop gone wobbly [Fig 5.34] ... and imagined it hit by even a barely energetic slow neutron. The neutron would add its energy to the whole. The nucleus would oscillate. In one of its many random modes of oscillation it might elongate. Since the strong force operates only over extremely short distances, the electric force repelling the two bulbs of an elongated drop would gain advantage. The two bulbs would push farther apart. A waist would form between them. The strong force would begin to regain the advantage within each of the two bulbs. It would work like surface tension to pull them into spheres. The electric repulsion would work at the same time to push the two separating spheres even farther apart.*

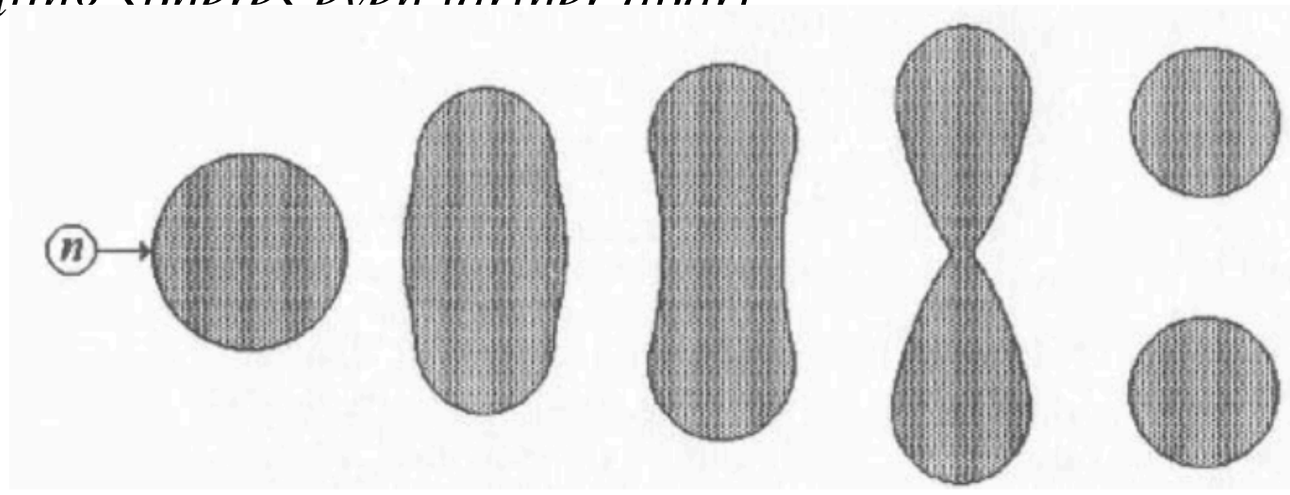


Fig 5.34 The stages in the fission of a drop of nuclear 'fluid' into two drops after the capture of a neutron.

Each of the two spheres would be a medium-sized nucleus; in the case of the fission of a uranium nucleus, $Z_U = 92$, they might well be a barium nucleus, $Z_{Ba} = 56$, and a krypton nucleus, $Z_{Kr} = 36$.

The energy released in the process arises from the intense electrostatic repulsion that acts between the middle-sized nuclei at the moment of their formation.

Most of this energy subsequently appears as the kinetic energy of these nuclei.

In order to undergo fission, a nucleus must first surmount (or tunnel through) the substantial energy barrier that corresponds, in the liquid-drop model of nuclear fission, to the shaping of the waist or constriction in the parent nucleus (Fig 5.35).

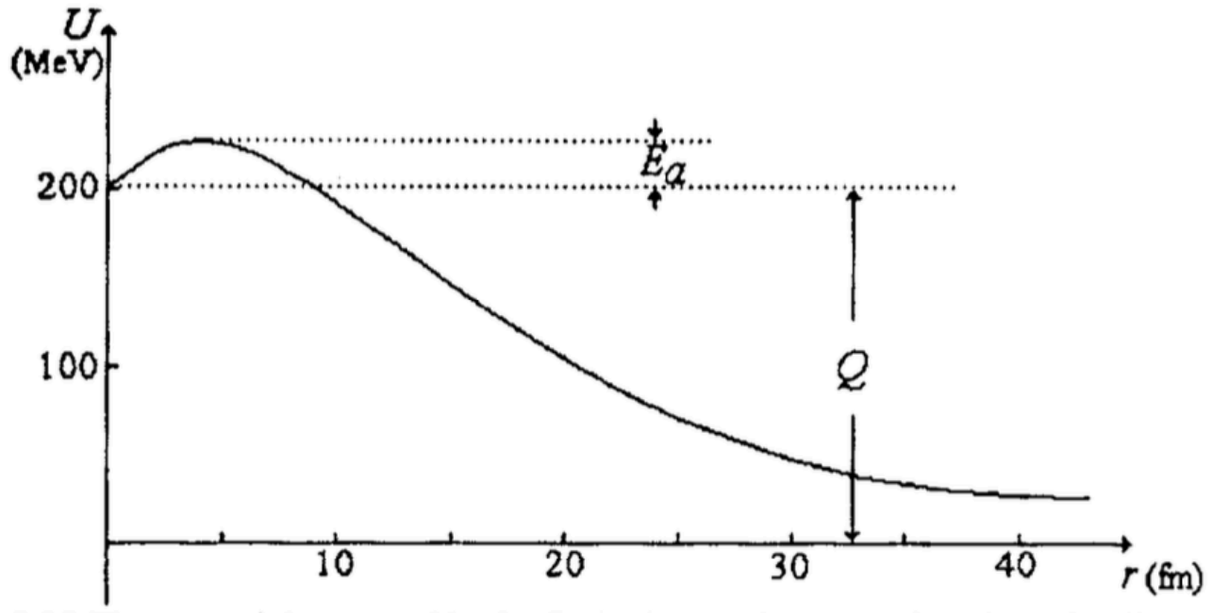


Fig 5.35 The potential energy, U, of a fissioning nucleus as a function of a distortion parameter, r, that measures its deviation from a spherical shape. The minimum energy required to initiate a reaction, E_a, is its threshold energy.

Thus, the spontaneous fission of heavy nuclei is a rare event.

For example, just one out of every two million U-238 nuclei decays by fission; all the rest decay by α emission.

The nucleus can obtain the fission threshold energy it requires in two ways: by the absorption of a high-energy photon or by the capture of a neutron.

In the latter process, the neutron initially adds cohesion to the nucleus without adding any electrostatic repulsion.

The extra cohesion adds to the binding energy of the nucleus.

The surplus energy thereby released provides the threshold energy required for fission to occur.

Since the pairing of nucleons reduces their energy, the total amount of energy provided by neutron capture depends, *inter alia*, on the number of neutrons in the nucleus.

The capture of a neutron by a nucleus containing an odd number of neutrons provides more nuclear energy than its capture by a nucleus containing an even number of neutrons.

In the former case, an unpaired neutron becomes paired; in the latter case, a lone unpaired neutron is added to a previously fully paired nucleus.

Thus, neutron capture induced fission should occur more readily in nuclei containing an odd number of neutrons.

This is borne out by the radiative capture and fission patterns of many heavy nuclei.

For example, the capture of a neutron by Th-232 and U-238, both of which contain an even number of neutrons, produces the compound nuclei Th-233 and U-239, respectively.

The incorporation of the neutron into their structure makes about 5MeV of energy available to these compound nuclei.

However, the threshold energy for their fission is about 6MeV.

Consequently, without additional energy, the probability of these nuclei undergoing fission is negligible.

Thus, irradiating Th-232 and U-238 with slow (thermal) neutrons results in radiative capture reactions, $^{232}Th(n,\gamma)^{233}Th$ and $^{238}U(n,\gamma)^{239}U$, respectively, and not fission.

The products of these reactions undergo - decay. In the case of U-239, this produces the relatively stable transuranic isotope Pu-239 ($T_{1/2}$ = 24,000years).

$$
\begin{aligned}
n + {}^{238}_{92}\mathrm{U} &\rightarrow {}^{239}_{92}\mathrm{U} + \gamma \\
{}^{239}_{92}\mathrm{U} &\rightarrow {}^{239}_{93}\mathrm{Np} + \beta^- + \bar{\nu} \qquad T_{1/2} = 24\,\mathrm{m} \\
{}^{239}_{93}\mathrm{Np} &\rightarrow {}^{239}_{94}\mathrm{Pu} + \beta^- + \bar{\nu} \qquad T_{1/2} = 2.3\,\mathrm{d}
\end{aligned}
\tag{5.55}
$$

Compound nuclei that do undergo fission can be obtained from Th-232 and U-238 by the capture of fast neutrons (>1 MeV); the kinetic energy of the fast neutrons provides the additional energy required for their fission.

For example, unless it is pre-empted by radiative capture, i.e., by the emission of a γ-ray from the excited compound nucleus, the incorporation of a 1.3MeV neutron into a U-238 nucleus generally results in its fission.

By contrast, the nuclei U-233, U-235 and Pu-239 all contain an odd number of neutrons.

The capture of a neutron by these nuclei produces the compound nuclei, U-234, U-236 and Pu-240, respectively.

Incorporating the neutron into their structure makes about 6MeV of energy available to these compound nuclei.

This is more than the threshold energy, ~5MeV, for their fission.

Thus, any neutron-slow (thermal) neutrons as well fast (high-energy) neutrons-can induce the fission of nuclei such as U-233, U-235 and Pu-239.

Unless pre-empted by rapid radiative capture, the intermediate compound nuclei undergo fission within an instant ($\sim 10^{-12}$s) of their formation without requiring any additional energy.

The values of the radiative capture and neutron induced fission cross-sections, $\sigma(n;\gamma)$ and $\sigma(n,f)$, listed in Table 5.18, reflect the probabilities of the two reactions occurring when the nuclei U-233, U-235, U-238 and Pu-239 are irradiated with fast or slow (thermal) neutrons.

Except in the case of the U-238 nuclei, the probability of fission is over a hundred fold greater when slow rather than fast neutrons are employed.

Table 5.18: The Radiative Capture and Fission Cross-sections of Some Fissionable (Fissile) Nuclei (in barns)

Nucleus	Thermal neutrons (~0.025eV)		Fast neutrons (>1MeV)
	Radiative Capture, $\sigma_{(n,\gamma)}$	Fission, $\sigma_{(n,f)}$	Fission, $\sigma_{(n,f)}$
U–233	47	530	-
U–235	95	586	1.22
U–238	2.7^{64}	negligible	0.70
Pu–239	270	752	1.73

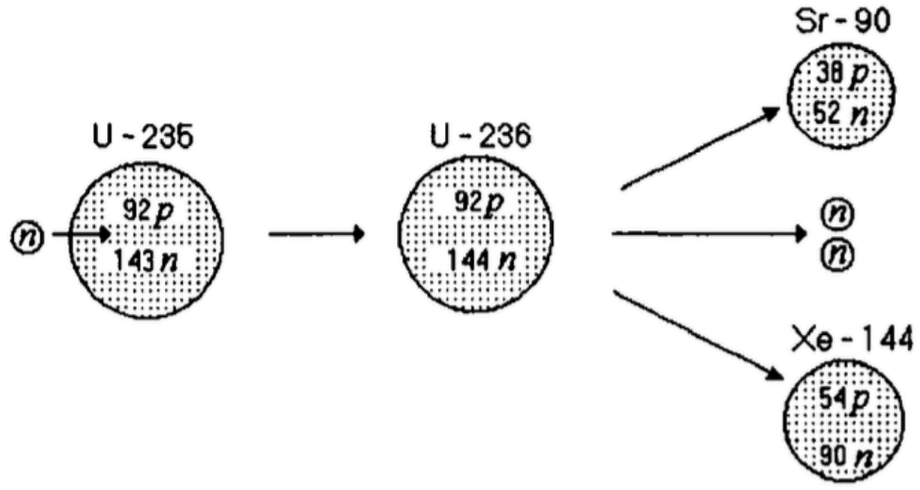
The fission of a compound nucleus is generally asymmetric, producing two different medium-sized nuclei. It is also accompanied, as we would expect, by the release of free neutrons.

A typical example of such a fission reaction is

$$n + {}^{235}_{92}\mathrm{U} \rightarrow {}^{236}_{92}\mathrm{U}^{*} \rightarrow {}^{144}_{54}\mathrm{Xe} + {}^{90}_{38}\mathrm{Sr} + 2n \qquad (5.56)$$

This reaction is shown schematically in Fig 5.36.

Fig 5.36 One of the possible modes of the neutron-capture induced fission of U-235. The radioactive isotopes Sr-90 and Xe-144 decay by β- emission that converts their excess neutrons into protons until stability is reached.



The fission products in this example, namely, the medium-sized isotopes Xe-144 and Sr-90, represent just one possible outcome.

Any pair of isotopes whose combined atomic number equals that of the fissioning nucleus can in principle be produced, though some pairs are found more often found than others.

The isotopes are initially in an excited nuclear state and are usually radioactive; they generally stabilize through a series of β- and γ decays.

Typically, the average binding energy per nucleon in the fission products is about 0.9MeV greater than that in the original nucleus.

Consequently, the energy released from the fission of a nucleus possessing 235 nucleons is about 235 x $0.9 \approx 200$ Mev.

This energy is divided between (i) the kinetic energy of the fission products, the isotopes and fast neutrons; (ii) the instantaneous radiation released by the process; (iii) the radioactivity of the isotopes, as shown in Table 5.19.

Table 5.19: The Average Disposition of the Energy Released from the Fission of U–235 in MeV

Form of the Energy	Disposition	Amount
Kinetic energy:	Isotopes	165
	Fast neutrons	5
Instantaneous radiation:	γ-rays	7
Radioactivity of the Isotopes:	β decay	23
Total		200

The fission of a U-235 nucleus produces on average 2.5 neutrons.

These are fast neutrons, with a mean kinetic energy is about 2 MeV and a mean velocity of ~10^7m/s.

5.4.7 Nuclear Chain Reactions

The notion that the considerable energy released by neutron induced nuclear fission could be usefully exploited was first contemplated by Szilard in 1934; he even filed a patent application for the process.

Much later, he related that the idea occurred to him as he was waiting to cross the road in London.

> *As the light changed to green and I crossed the street, it suddenly occurred to me that if we could find an element which is split by neutrons and which would emit two neutrons when it absorbs one neutron, such an element, if assembled in sufficiently large mass, could sustain a nuclear chain reaction.... In certain circumstances it might be possible to set up a nuclear chain reaction, liberate energy on an industrial scale and construct atomic bombs.*

Shortly after the discovery of the neutron induced fission of uranium by Hahn and Strassman, the surplus of fast neutrons released in the process was also detected.

The element that emits two (or more) neutrons when split by one neutron had been found; it was uranium.

Szilard's notion of a nuclear chain reaction had become a reality.

The two and a half neutrons released, on average, from the fission of a uranium nucleus could in turn induce the fission of other nuclei, and those released from these other nuclei could go on to induce the fission of still further nuclei.

Thus the chain reaction would be propagated with each extra fission adding its ~200MeV of liberated energy to the total.

A nuclear chain reaction will only be self-sustaining if, on average, at least one of the neutrons released from each nuclear fission induces the fission of another nucleus.

Or, equivalently, if at the beginning of each fission generation, at least the same number of neutrons are present in the system as were there at the start of the previous generation.

This provision is expressed mathematically in terms of a parameter. k. called the multiplication factor.

This is defined as the number of the neutrons released from the fission of each nucleus that, on average, induce the fission of another nucleus.

If $k < 1$, on average, too few neutrons induce further fissions and so the chain reaction stops. This is called the *subcritical* state.

If $k = 1$, on average, at least one of the neutrons released from the fission of each nucleus induces the fission of another nucleus. The chain reaction will be self-sustaining, generating energy at a steady rate. This is called the *critical* state.

If $k > 1$, on average, more neutrons induce further fissions than are absorbed or lost. The rate at which the process generates energy will grow exponentially, terminating only when the system has exhausted all its fissile resources. This is called the *supercritical* state (Fig 5.37).

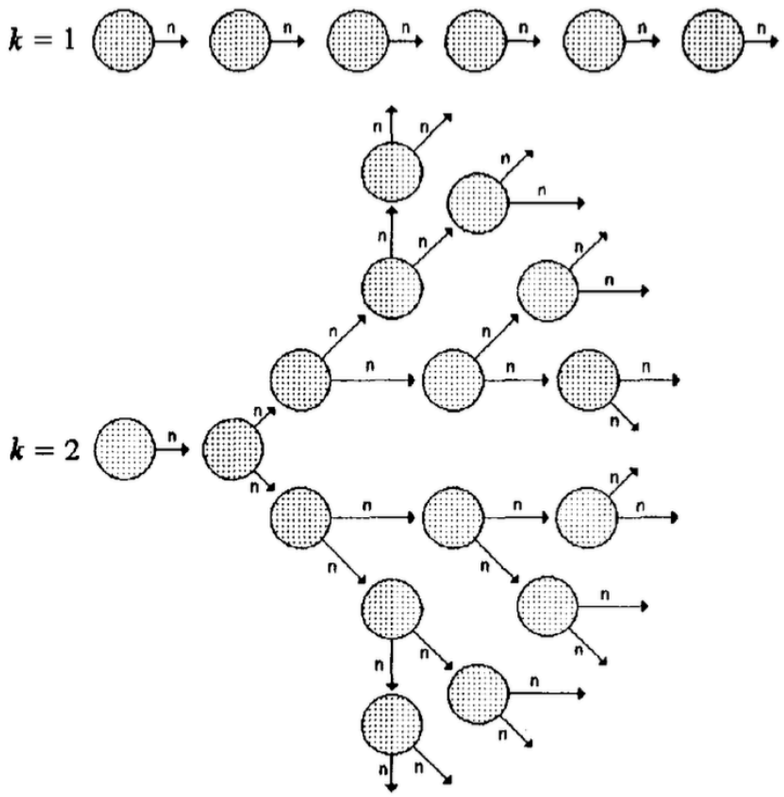


Fig 5.37 Nuclear fission chain reactions. In the critical state, k = 1, just one neutron from each fission induces another neutron. In the supercritical state shown, k = 2, two neutrons from each fission induce further fissions.

The multiplication factor of an infinitely large mass of fissile material, one from which effectively no neutrons can escape, is designated k_∞ .

For pure U-235, the multiplication factor $k_\infty \approx 2$; for a mixture comprising 50% U-235 nuclei and 50% inert non-fissile nuclei, $k_\infty \approx 1$, and for a mixture containing less than 50% U-235, $k_\infty < 1$.

The multiplication factor of a finite body of fissile material will always be less than k_∞.

How much less depends on the ratio between the number of neutrons that collide with the nuclei in the body of fissile material and the number that escape from its surface.

This ratio, in turn, depends on the body's shape and dimensions.

In general, the smaller the body the higher the proportion of the neutrons that escape and the lower is the value of k_∞.

The mass of a body or formation of fissile material for which $k_\infty = 1$ is called its critical mass.

A body or formation of fissile material larger than the critical mass will be *supercritical*; one smaller than the critical mass will be *subcritical*.

Whether the initial fission in the chain reaction is induced by a fast or slow neutron, all the neutrons released in subsequent fissions are fast neutrons.

Thus, unless special arrangements are incorporated to slow the neutrons down, a chain reaction in a supercritical chunk of fissile material will be overwhelmingly a fast neutron reaction.

Once initiated, the chain reaction will propagate rapidly, terminating only when all the fissile resources have been exhausted or after the chunk has blown apart, whichever comes first, at which point the system becomes sub-critical.

Since there are always some free neutrons around to initiate the reaction, for instance from cosmic rays, any supercritical body of fissile material is potentially explosive.

In order to prevent such eventualities, fissile material must be kept in small pieces each of whose mass is less than the critical mass.

Given that σ_t is the cross-section for a particular chain of neutron induced fission reactions, the radius of a spherical critical mass of highly fissile material containing N fissile nuclei per unit volume is approximately equal to the probable distance

$$d = 1/N\sigma_t \tag{5.57}$$

that a neutron must travel in the material, during a chain reaction, before inducing fission.

With a density, ρ_U= 19,000 kg/m^3, the number of atoms, N, in a unit volume of solid U-235 is

$$N = \frac{N_{Avogadro} \cdot \rho_U}{0.235} = \frac{6.0 \cdot 10^{23} \cdot 19{,}000}{0.235} \approx 4.85 \cdot 10^{28}\,\mathrm{m}^{-3} \tag{5.58}$$

The cross-section for a chain of neutron induced fission reactions is equal to twice the fission cross-section: $\sigma_t \approx 2\,\sigma_{nf}$.

The fast neutron fission cross-section of U-235 is $\sigma_{nf} = 1.22$b. Thus, the radius of a critical sphere of pure U-235 is

$$d = \frac{1}{N\sigma_t} = \frac{1}{4.85 \cdot 10^{28} \cdot 2 \cdot 1.22 \cdot 10^{-28}} = 0.085\text{m} \tag{5.59}$$

The mass of a sphere of pure U-235 of radius 0.085m is about 50kg.

This is the critical mass of a sphere of pure U-235; a larger sphere would detonate the moment it was assembled.

The critical mass of a sphere of Pu-239 is about a quarter as large, ~l0 kg.

Naturally occurring uranium, which predominantly comprises the two isotopes, U-238 (99.3%) and U-235 (0.7%), can be safely kept in ingots of any size; any quantity of it, no matter how large, is sub-critical.

This is because of the high threshold energy, ~6MeV, for neutron induced U-238 fission. Of this total, about l MeV must be supplied by the neutron's kinetic energy.

However, although their average kinetic energy is ~2 MeV, only three-quarters of the fast neutrons released by fissioning uranium nuclei actually have energies greater than l MeV and only a quarter of these induce a fission before being slowed to energies below l MeV.

Thus the multiplication factor, k_∞, for an infinite mass of pure U-238 has a maximum value of

$$k_\infty = 2.5 \cdot 0.75 \cdot 0.25 = 0.47 \tag{5.60}$$

With a concentration of just 0.7%, the more fissile U-235 isotope cannot make up the difference required for the material to sustain a fission chain reaction and so naturally occurring uranium is safe against fast neutron fission.

Only if the U-235 concentration was increased more than tenfold, would an infinite mass of the mixture become critical.

Making such 'enriched' uranium requires separating the isotopes in the naturally occurring material.

This can be done by mass spectography, or, more efficiently, by first converting the element into the gaseous compound uranium hexafluoride, $^{235}UF_6$ and $^{238}UF_6$, and then segregating the molecules comprising the respective isotopes by membrane diffusion or centrifuging.

In order to make nuclear fission both sustainable and controllable, the neutrons emitted from the fissioning nuclei must first be slowed down.

This requires the incorporation of a neutron-slowing substance such as paraffin wax into the body of fissile material.

The incorporation of such a moderator into the fission process is a key feature of the devices used for the controlled release of nuclear energy known as atomic piles or atomic reactors.

5.4.8 Fission by Fast Neutrons - Bombs

Within a few months of the discovery of nuclear fission, the governments of all the major world powers had been advised by their scientists of its possible military applications.

> *Work began first in Germany, where the Reich Ministry of Education convened a secret conference on April 29th, 1939, that led to a research program and a ban on uranium exports. Independently, during the same month, a Japanese army general ordered military applications explored. In the United States, Hungarian emigre physicists Leo Szilard, Edward Teller and Eugene Wigner communicated their concerns about German developments to President Franklin D. Roosevelt through a letter from Albert Einstein that Roosevelt reviewed on October 11, 1939. The British took a first look in 1939, stalled, and then, catalyzed by a memorandum from the physicists Otto Frisch and Rudolph Peierls, emigres from Nazi Germany, began in earnest early in 1940. The Soviet physicist Igor Kurchatov73 alerted his government to the possible military significance of nuclear fission in 1939.*

All of the governments were informed of the possibility of constructing a 'super-bomb' the blast from which would destroy life in a wide area: "The size of this area is difficult to estimate, but it will probably cover the center of a big city... as a weapon, the super-bomb would be practically irresistible",.

Explosive power, i.e., the intensity of the shock (pressure) wave produced by an explosion, depends not only on the amount of energy released but also on the rate at which it is released.

In an efficient explosive device, as much energy as possible must be released in as short a period of time as possible.

The heat released by the exploding material raises its temperature and this, in turn, raises the pressure inside the device.

It is the sudden release of this pressure as the device bursts that generates the destructive shock wave of an explosion.

The longer it is delayed, the stronger the explosion.

For this reason, conventional explosives are often encased in heavy materials-tampers-whose inertia confines the blast for a few more microseconds, delaying the release of the pressure.

Taking the typical muzzle velocity of a rifle, ~ l000m/s, as indicative of the average velocity with which the fragments separate, the distance between them will be ~l cm just l0μs after detonation.

The detonation of a fission bomb requires the almost instantaneous assembly of a supercritical body of fissile material; the presence of just a single free neutron in this body will then initiate a nuclear fission chain reaction.

Once initiated, the chain reaction in the assembly must proceed both fast enough and long enough to produce a powerful explosive blast.

Expansion of the fissile material by just a few centimeters will halt the chain reaction; the assembly will no longer be supercritical.

Thus, the chain reaction must be completed within one or two microseconds.

At a typical velocity of 10^7m/s, a fast neutron will cover the probable distance it must travel between fissions in pure U-235, viz. 0.085m, in about 0.01μs.

This then is the typical lifetime of a fast fission generation.

Slowing down the neutrons requires about a microsecond for each generation, i.e., 100,000 times the lifetime of each fast fission generation.

This is far too slow for an explosion to develop.

Thus, even though the fission cross-sections of U-235 and Pu-239 for slow neutrons are some 300 times those for fast neutrons, atomic bombs employ fast neutron fission.

A critical mass of fissile material, viz. about 50kg of pure untampered U-235 or 10kg of pure untampered Pu-239, is the minimum quantity required for a fission bomb.

Since the mass begins to disintegrate and to fly apart once the reaction starts, the fission of the entire critical mass is unrealizable.

Typically, just one kilogram, ~4 moles, of the fissile material undergoes fission; an efficiency of just 2% for a 50kg mass of U-235.

Under the conditions prevailing in an explosion, the multiplication factor $k \approx 1.5$.

Hence, the number of generations, ν, required for the fission of 4 moles of fissile material is given by the sum:

$$4 \cdot 6 \cdot 10^{23} = 1 + k + k^2 + \ldots + k^{\nu-1} = \frac{\left(1 - 1.5^{\nu}\right)}{1 - 1.5} \tag{5.61}$$

from which $\nu \approx 1\ 130$.

Given, that the lifetime of a fast fission generation is typically just 0.01μs, the fission of the one kilogram of U-235 is completed in under 1.5μs.

Thus, the reaction is all over before the device has had time to expand to any significant degree.

Assuming that the fission of each uranium nucleus releases 165MeV of mechanical (kinetic) energy, the total blast energy released by the fission of one kilogram (~ 4 moles) of U-235 will be about $4\text{-}6\text{-}10^{23}$ x165 $\approx 4\text{x}10^{26}$MeV: The energy released in the explosion of a ton of conventional chemical explosive, TNT or dynamite, is about $4\text{x}10^{9}$ J.

Thus the fast fission of just one kilogram of U-235 is equivalent to the detonation of

$$\frac{4 \cdot 10^{26} \cdot 1.6 \cdot 10^{-13}}{4 \cdot 10^{9}} \approx 16{,}000 \text{ tons} \tag{5.62}$$

(16 kilotons) of conventional chemical explosive.

The efficiency of the device can be enhanced by encasing the fissile material in a heavy neutron-reflecting tamper.

This has the double effect of slowing the expansion of the fissioning material and reducing the neutron loss from its surface.

The tampers are typically made of beryllium that reflects neutrons back into the fissioning mass and natural uranium metal which is one of the densest of metals.

Their overall effect is to reduce the critical mass.

Two different methods have been employed to assemble the supercritical body of fissile material required for the efficient detonation of a fission bomb.

The first is the rapid fusion of several sub.critical masses of fissile material into a single supercritical mass; the second is the rapid compression of a single sub.critical mass of fissile material into a supercritical body of denser material.

Compressing the fissile material increases its density, which in turn reduces the probable distance a neutron must travel in the material before it induces fission.

Thus, in the compressed state, the fissile material has a smaller critical mass; doubling the density of the fissile material results in a four-fold reduction in the critical mass.

The first method was employed in the bomb nicknamed 'Little Boy' that destroyed Hiroshima on August 6th 1945.

In this device, a subcritical bullet of almost pure U-235 was fired into subcritical target forming a tampered supercritical assembly.

In order to ensure the efficient detonation of the bomb the moment the supercritical assembly was created, an initiating neutron source, a small amount of an a emitter such as polonium coated with a light element like lithium or beryllium, was incorporated into the bomb (Fig 5.38).

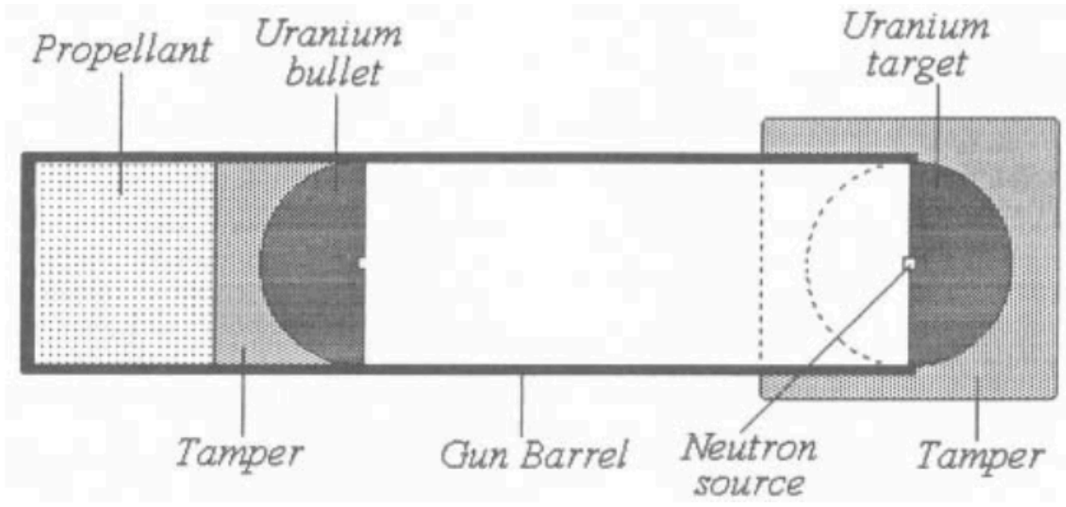


Fig 5.38 Schematic diagram of the U-235 bomb 'Little Boy'. Igniting the propellant fires the sub-critical bullet at the sub-critical target. Together they form a supercritical assembly. The neutron source ensures the bomb's efficient detonation.

The assembly of a supercritical mass of Pu-239 from a number of subcritical pieces will not work; plutonium is so reactive that the device will predetonate, i.e., the fission reaction will start too soon, "melting the bullet and target before the parts had time to join to produce a high-yield explosion".

Thus the second method of detonating a fission bomb, compressing the fissile material to increase its density, must be used in plutonium based bombs.

The bomb, nicknamed 'Fat Man', that devastated Nagasaki on August 9th 1945 was a plutonium bomb.

Although a solid 5kg tampered sphere of Pu-239 would explode immediately on assembly, a hollow thick-walled spherical shell of the same mass is safely subcritical.

The compression and inward collapse of the hollow shell would raise its density and at a sufficiently high density it would become supercritical.

In bombs of the 'Fat Man' type, a plutonium shell is enveloped in an intricate assembly of different types of conventional explosives.

The timed detonation of these explosives produces an implosive shock-wave that is focused on the centre of the shell and crushes the plutonium into a dense supercritical mass (Fig 5.39).

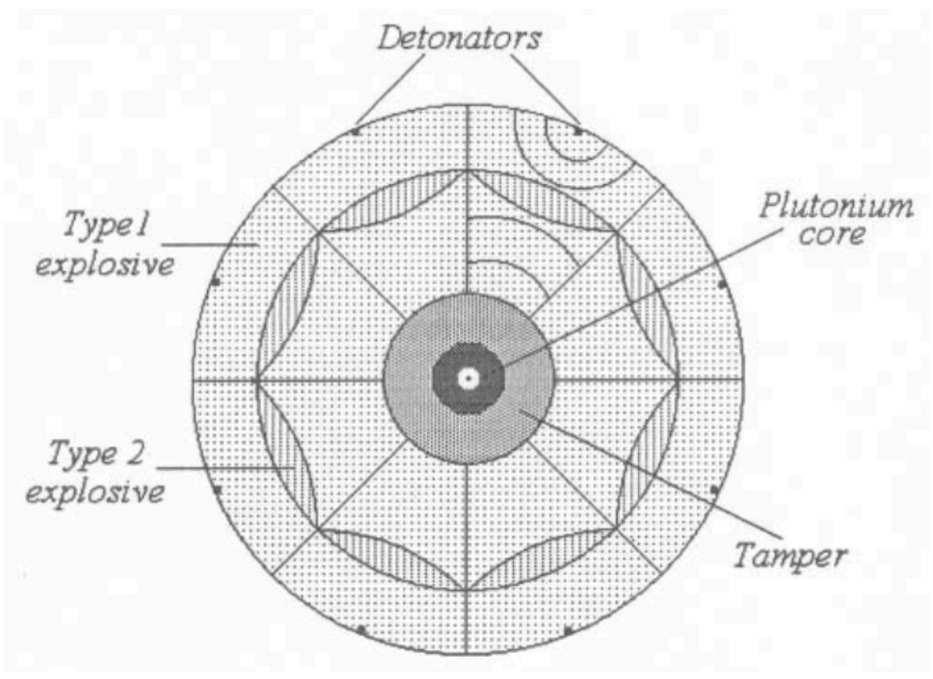


Fig 5.39 The convex shock wave generated by the detonation of the outer layer of the fast burning type I explosive is focused by the lenses of the slower burning type 2 explosive. The intensity of the concave wave that emerges from the lenses is reinforced by the inner layer of type 1 explosive. It is smoothed out by the heavy tamper and then crushes the plutonium core into a supercritical state.

The plutonium isotope, Pu-239, used in the 'Fat Man' bomb was manufactured by irradiating natural uranium with neutrons.

The radiative capture reaction, $^{238}U(n,\gamma)^{239}U$, between the uranium's preponderant isotope, U-238, and the neutrons is followed by two β- decays. during its exposure also reacts with the neutrons.

These raise the atomic number by two, thus producing the plutonium-equation (5.55).

The plutonium was extracted chemically from the irradiated material.

The isotope Pu-239 that accumulates in the irradiated material.

This radiative capture reaction, $^{239}Pu(n,\gamma)^{240}Pu$, produces the isotope Pu-240.

The longer the material is irradiated, the more Pu-240 accumulates in it.

Although the isotope Pu-240 does not undergo neutron induced fission, it has a small but not insignificant probability of undergoing spontaneous fission, with the release of a flood of neutrons.

Thus, the presence of this isotope in the plutonium used in a bomb could cause it to go off too soon.

Imploding at a speed of ~1000m/s, the plutonium shell takes about 4μs to achieve the desired density.

The release of a neutron during this time from the spontaneous fission of a Pu- 240 nucleus would cause the core to pre-detonate and 'fizzle' instead of exploding.

Thus, the plutonium extracted from uranium that has been irradiated by neutrons for more than a few weeks is generally too contaminated with Pu-240 to be considered weapons grade material.

For a few seconds after detonation, the temperature of the fireball produced from a nuclear explosion exceeds that of the surface of the sun; any object exposed to the fireball receives an almost instantaneous intense dose of thermal radiation.

A 20-kiloton explosion delivers enough thermal radiation to ignite combustible material at a distance of 2000m.

About a third of the fatalities at Hiroshima and Nagasaki were due to flash bums.

Since the bombs were detonated at an altitude of about, 2000 feet, blast damage was maximized.

However, although radioactive material with an activity of ~10^6Ci was generated in the blast, a month afterwards virtually no trace of radioactive fall-out was found in Hiroshima or Nagasaki.

It appears, that after the initial burst of nuclear radiation, an air-blast leaves little radioactive debris in the vicinity of the explosion.

Both cities were subsequently rebuilt on the same sites.

More modern nuclear weapons employ both fission and fusion; In a typical example, the instantaneous burst of γ rays released by the detonation of a primary fission device initiates more powerful fission and fusion reactions in a secondary nuclear assembly.

Because the γ rays travel at the speed of light, they arrive at the secondary assembly well ahead of the shock wave generated by the primary explosion; this provides time for the secondary assembly to react before being blown apart.

The secondary assembly comprises layers of thermonuclear material, lithium-6 deuteride ($^6Li^2H$), and fissile material, Pu-239, that are encased in a U-238 tamper and surrounded by configured layers of dense plastic foam (Fig 5.40).

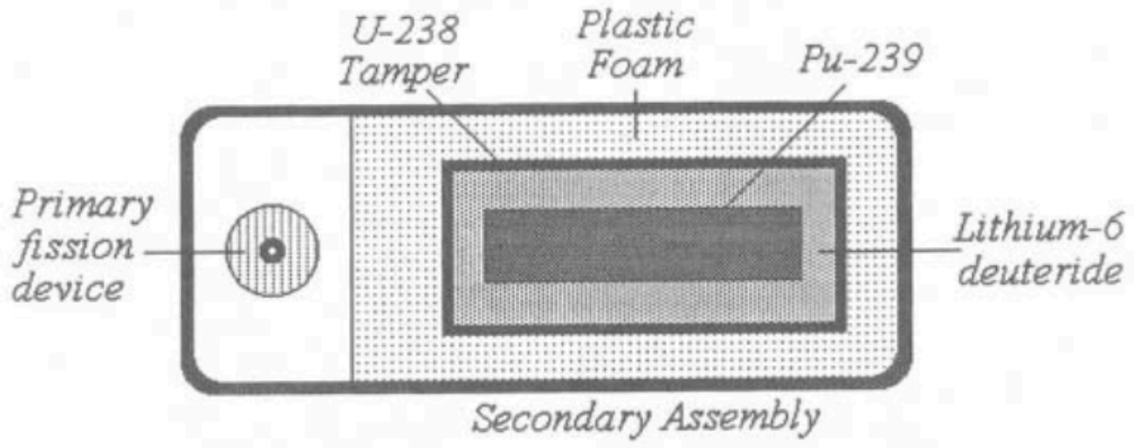


Fig 5.40 Schematic diagram of a fission-fusion-fission nuclear bomb. The γ rays released from the primary fission device vaporize the plastic foam. This compresses the core of fissile and thermonuclear materials initiating both fusion and fission reactions. These rapid reactions are sustained by the heat and neutrons they release until the bomb blows itself apart.

The first reaction that occurs in the secondary assembly is the absorption of the y ray energy from the primary explosion by the shaped pieces of foam.

This causes their instantaneous vaporization and conversion into a rapidly expanding plasma.

The pressure wave generated by the plasma compresses and implodes the fissile and thermonuclear materials.

The now supercritical layers of fissile Pu-239 react, releasing a flood of neutrons and still more energy.

Neutrons escaping into the thermonuclear material from the chain reaction in the Pu-239 react with the Li-6, releasing tritium, ^3H, and helium, ^4He, nuclei and yet more energy:

$$^6_3\mathrm{Li} + n \rightarrow {}^4_2\mathrm{He} + {}^3_1\mathrm{H} \quad Q = 4.9\mathrm{MeV} \tag{5.63}$$

At the tremendous temperatures and pressures by now prevailing in the secondary assembly, the tritium produced from this reaction fuses with the deuteron in the thermonuclear material to produce still more neutrons and even more energy:

$$ {}^{3}_{1}\mathrm{Li} + {}^{2}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + n \quad Q = 17.6\mathrm{MeV} \tag{5.64} $$

Any surplus neutrons remaining in the device at this stage will possess energies in excess of the threshold energy required to induce the fission of the U-238 in the tamper and so this too will start to react, adding its contribution to the aggregation of energy and neutrons.

All these reactions take place in parallel and at great speed until the bomb finally bursts.

The total amount of energy released depends on the amount of thermonuclear and fissile material packed into the secondary assembly.

Thermonuclear weapons have TNT equivalencies between 100,000 tons (a hundred kilotons) and 10,000,000 (ten megatons) depending on the target for which they are intended.

At a distance r from the point of detonation, the maximum value, p_{max}, of the pressure in the shock wave from a powerful explosion of energy E is approximately given by:

$$ p_{\max} = E/r^3 \tag{5.65} $$

Given that the destructive radius of a bomb containing 0.25tons of TNT is about 50m, that of a one megaton bomb is about $\sqrt{4 \cdot 1,000,000}$ x 50 ≈ 8000m .

This is wide enough to encompass even the largest cities in the world.

It is a sobering thought that even now, long after the end of the Cold war, there are still some 20,000 nuclear weapons of various types and capacities in the arsenals of the world's armies ready for use at a moment's notice.

5.4.9 Fission by Slow Neutrons - Nuclear Reactors

The first atomic reactor was constructed by a team led by Fermi in 1942. Its successful operation established two important points:

1. **Nuclear fission could be controlled and so could be used as a useful energy source;**
2. **The fissile artificial isotope Pu–239, previously prepared only in microgram quantities by irradiating uranium with the relatively weak neutron flux produced from a cyclotron, could now be manufactured in industrial quantities.**

Fermi's reactor was fueled with natural uranium.

Detailed calculations show, although natural uranium, whose composition is 99.3% U-238 and 0.7% U-235, is not susceptible to fast fission, because of the very large slow neutron fission cross-section of U-235, a sufficiently large mass of natural uranium can sustain a slow (thermal) neutron fission chain reaction,

However, whether the initial fission is induced by a fast or slow neutron, all the neutrons released in subsequent fissions will be fast neutrons.

Thus, sustaining a chain reaction in natural uranium requires slowing down-thermalizing-the neutrons emitted from each fissioning uranium nucleus before they collide with another nucleus.

This is achieved by embedding the natural uranium in a 'moderating' material like paraffin wax that absorbs the kinetic energy of the neutrons.

The choice of moderating material involves striking a balance between two factors: the material's efficiency in slowing neutrons versus its nuclear cross-section for neutron capture.

Whereas the former should be as high as possible, the latter should be as small as possible.

The presence in the assembly of a material that readily captures neutrons is tantamount to 'poisoning' the chain reaction; neutrons that should be available for propagating the chain reaction would instead be captured by the material.

In addition, the whole assembly must be made very large in order to prevent the loss of neutrons as a result of their leakage from its surface.

In elastic collisions between neutrons and stationary nuclei of mass number A, the energy of the incident neutrons is reduced on average by about A/(l+ A) in each collision.

Thus, from an initial energy E_0, the energy, E, of a neutron will have been reduced to about

$$E = E_0\left(A/(1+A)\right)^n \tag{5.66}$$

after n collisions.

Clearly, the smaller the mass number of the stationary nucleus, the less collisions will be required to thermalize fast neutrons, i.e., to reduce their average energy from 2MeV to 0.025eV.

This suggests that the best moderating material to use would be hydrogen nuclei, for which A = 1.

According to standard calculations just 23 collisions between the neutrons and the hydrogen nuclei would be required to achieve the required reduction in neutron energy.

However, hydrogen nuclei have an appreciable cross-section, $\sigma = 0.31$b, for absorbing slow neutrons.

Thus, the presence of hydrogen in the moderating material would cause a significant loss in the number of neutrons available for fission.

The second best choice is deuteron, ^{2}H.

Thermalizing the 2MeV neutrons would require more collisions between the neutrons and the deuterons than with hydrogen.

However, deuteron's cross-section for slow neutron capture is only 0.00065b.

The common isotope of oxygen, ^{16}O, also has a negligible cross-section for neutron capture.

Consequently, heavy water, $^{2}H_2{}^{16}0$, is a particularly good neutron moderator.

Whereas, on average, a slow neutron will be absorbed in ordinary water after just 150 collisions, in heavy water it will only be absorbed after 10,000 collisions.

Ordinary water can only be used as a moderator in reactors fueled by enriched uranium, i.e., uranium in which the concentration of the U-235 is 2.5% or more.

Since, at the time, heavy water was unavailable in sufficient quantities, Fermi had to use the next easily available small nucleus with a negligible cross-section for neutron capture; this was the carbon-12 nucleus.

Over a hundred collisions would be needed to thermalize each neutron but few would be lost as they could collide on average a thousand times before being absorbed by a ^{12}C nucleus.

In the reactor Fermi constructed, alternate layers of solid blocks of very pure graphite and of blocks whose cores had been drilled out and loaded with natural uranium were piled up on one another; the assembly was called an atomic pile.

Calculations showed that a sufficiently large pile would be supercritical.

In order to prevent the pile from becoming critical, or even supercritical, before its assembly was completed, removable cadmium rods were incorporated into its design.

Natural cadmium has a remarkably high cross-section, u 2000b, for the radiative capture of thermal neutrons.

Thus, so long as the cadmium rods remain inside the pile, it could never reach criticality, whatever its size.

Any neutrons released inside the pile would be absorbed by the cadmium before they could induce fission.

Removing the cadmium rods from an inherently supercritical reactor, would start off the fission chain reaction at once (Fig 5.41).

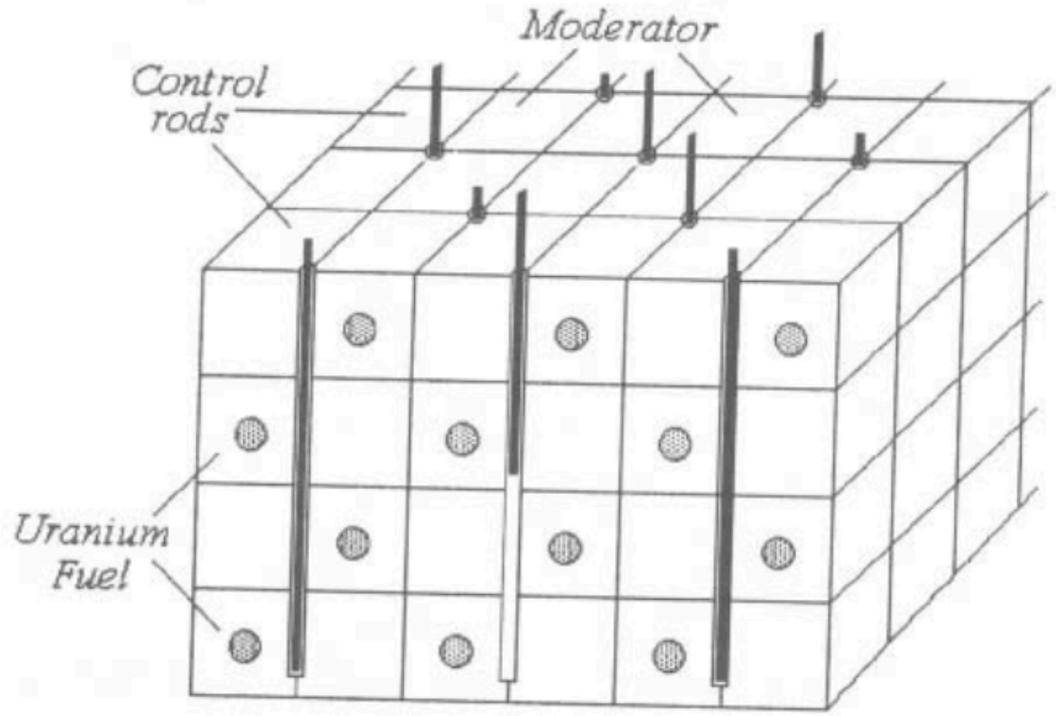


Fig 5.41 Diagram of a typical reactor core. Horizontal rods of natural or enriched uranium are packed into the moderating material. The flux of neutrons in the reactor and hence the rate of the fission reaction is controlled by raising or lowering the vertical neutron absorbing control rods.

Controlling or halting the chain reaction in the pile after it has become supercritical appeared, at first, to present an insurmountable problem.

Since the pieces of uranium are some 10-20cm apart in the reactor and the mean velocity of thermal neutrons is ~l000m/s, the average time spent by the neutrons in traveling between fissions is about l00μs.

Thus, a thousand generations can pass in just a tenth of a second.

Suppose, that a reactor in a critical state, k = 1, becomes just barely supercritical, k = 1.001.

Whereas in the critical state, the rate at which the reactor produces energy is constant, in the supercritical state its energy output grows exponentially.

Every tenth of a second, its energy output grows by $1.001^{1000} = 2.7$, and after just one second it will be almost 22,000 times greater than it was in the critical state.

No mechanical control can react as quickly as the nuclear chain reaction propagates; the cadmium rods could not be re-inserted into the reactor in time.

Thus, once it becomes supercritical, it would appear that a reactor will rapidly and unavoidably run out of control and ‘melt-down'.

Fortuitously, there is a factor that makes reactor control possible; it is the delay in the release of a small number of the neutrons.

On average, 0.6% of the neutrons are released only after a delay that can be as long as 30 seconds after the actual fission.

These ‘delayed’ neutrons are released by the unstable fission fragments, the radioactive medium-sized nuclei into which the fissile nucleus splits.

In the estimate we made of the average time interval between fissions, we assumed, wrongly, that all the neutrons are released promptly; we only took into account the motion of the slowed neutrons from one piece of uranium in the reactor to another.

Allowing for the delayed neutrons, the average time interval between fissions in a reactor is about one second, a thousand times longer than we had estimated.

Consequently, changes in a reactor's power output take place only very slowly; starting up a reactor from scratch can take hours.

Although their release is delayed, these neutrons contribute to the overall neutron flux in the reactor; without them it would not become critical.

Consequently, so long as a reactor is inherently subcritical without them, i.e., $k_{reactor} < 1.006$, its power production is controllable.

This sets an upper-limit on the rate at which nuclear reactors may be safely operated.

The reactor assembled by Fermi produced energy at a rate of just 200W; modern commercial reactors used for the production of electricity typically operate at power-ratings of thousands of megawatts.

Nuclear reactors that employ slow neutron fission are called thermal reactors.

The simplest form of commercial thermal reactor is the Pressurized-Water Reactor.

These reactors are fueled by slightly enriched uranium, 3% U-235; the water serves both as the coolant and moderator.

The reactor core is situated inside a pressurized vessel, which is itself enclosed inside a strong containment building.

Superheated water, at a pressure of about 150 atmospheres, enters the pressurized vessel at a temperature of about 280°C, heats up further as it flows through the reactor core and leaves at a temperature of about 310°C.

From here it passes into a heat exchanger, where it heats the water in a secondary circuit producing steam.

This in tum spins turbines which drive electric generators (Fig 5.42).

A typical commercial reactor operates at an overall power of 3000MW and generates electricity at a rate of 900MW.

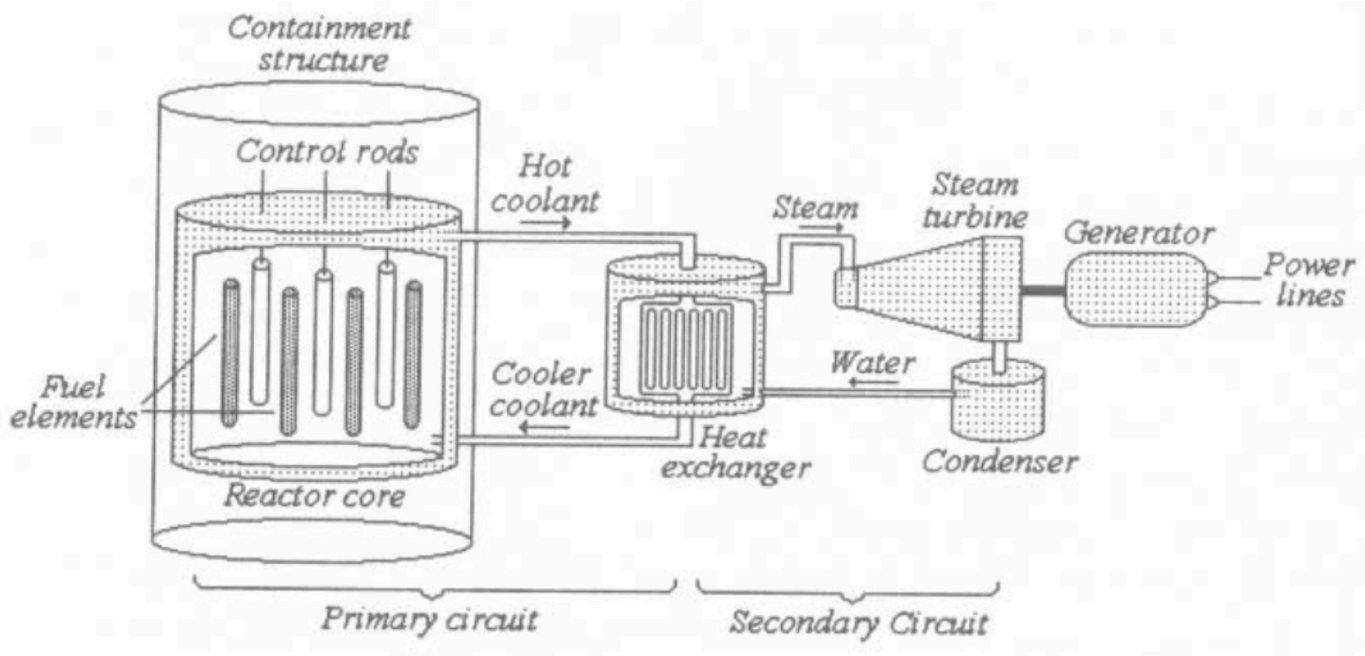


Fig 5.42 Schematic diagram of a typical nuclear power station. If the reactor is of the pressurized-water type, the fuel is enriched uranium and the coolant is superheated water; the water also acts as th moderator. In other reactor designs, natural uranium i the fuel, graphite is used as the moderator and the coolant may be steam, helium or carbon dioxide gas or liquid sodium.

The slow-neutron fission of its U-235 is not the only nuclear reaction that takes place in the reactor.

Surplus neutrons released from the fission of the U-235 nuclei react with the more plentiful U-238 nuclei also present in the fuel elements, resulting in its conversion into Pu-239 by the series of reactions listed in equation (5.53).

The artificial transuranic isotope Pu-239 is even more susceptible to slow-neutron fission than is U-235; its nuclear cross-section for slow-neutron fission is 752b.

Thus, as the percentage of U-235 in the fuel elements depletes, its role in the reactor is assumed by the fissile Pu-239.

After a few months more energy is being produced by the fission of the plutonium than the uranium.

Thus, a freshly fueled reactor can operate continuously for a year or more before having to be shut-down for maintenance and re-fueling.

Reactors that manufacture more fissile material than they consume are called breeder reactors.

By using highly enriched uranium or plutonium as the fuel, reactors can be constructed that dispense with the need for a moderator.

These reactors work by fast neutron fission and are called fast reactors to distinguish them from the thermal reactors that operate with slow (thermal) neutrons.

In other reactors called fast-breeders the core of enriched nuclear fuel is enveloped in a blanket of U-238; under suitable conditions, radiative capture converts the U-238 to Pu-239, producing more of it than was required to enrich the fuel in the core.

Although the fissile isotope Pu-239 is a by-product of every uranium fueled reactor, unless it is removed from the reactor within days or weeks of its manufacture it is usually so contaminated with Pu-240 that it is unusable for bomb production.

In almost all reactors designed primarily for electricity production, the removal of the plutonium requires closing down the reactor.

Since, for reasons of cost and service, this is done only at the longest possible intervals, by which time there is considerable Pu-240 contamination, the plutonium produced by such reactors is not usually of weapons grade.

Nevertheless, in their attempts to control the proliferation of fissile material and nuclear weapons, the nation states possessing the technology required for the construction of commercial reactors demand the return of the spent fuel elements containing the plutonium as a condition for supplying the fresh fuel elements without which the reactor will no longer work.

The isotope Pu-240 is not the only non-fissile substance that accumulates in a reactor's fuel elements during its operation; radioactive fission fragments, unstable medium-sized nuclei, also accumulate in its fuel elements.

One of these nuclei is the iodine isotope 1-135 which decays via Xe-135 into Cs-135:

$$\begin{aligned} {}^{135}_{53}\mathrm{I} &\rightarrow {}^{135}_{54}\mathrm{Xe} + \beta^- + \bar{\nu} \qquad T_{1/2} = 6.7\,\mathrm{h} \\ {}^{135}_{54}\mathrm{Xe} &\rightarrow {}^{135}_{55}\mathrm{Cs} + \beta^- + \bar{\nu} \qquad T_{1/2} = 9.2\,\mathrm{h} \end{aligned} \tag{5.67}$$

The isotope Xe-135 has an enormous radiative capture nuclear cross-section, $\sigma_{(n,\gamma)} = 2.6\text{x}10^6$ b. Its production in the reactor slows the chain reaction; its accumulation in the reactor would 'poison' it completely.

Because of its enormous appetite for neutrons, no Xe-135 will accumulate in a reactor as long as it is in operation and there is a substantial neutron flux through it; any Xe-135 produced from the decay of the 1-135 will be immediately swept up and removed by the radiative capture reaction, $^{135}\mathrm{Xe}(n,\gamma)^{136}\mathrm{Xe}$.

However, if the reactor is shut down, the Xe-135 does start to accumulate; there are no longer neutrons around with which it can react.

The Xe-135 reaches its maximum concentration in the reactor about 10 hours after the shut-down; it then starts to fall as the stock of I-135 declines.

Thus, after shutting down or drastically reducing the operation of a reactor, it cannot be immediately restarted. It must remain idle until the accumulated Xe-135 has decayed into Cs-135.

5.4.10 Nuclear Engineering: The Chernobyl Catastrophe

The experience of the past generation has shown that well-designed commercial reactors, in the hands of competent operators, are safe.

Over 1200 atomic reactors of various types are presently in operation throughout the world.

About a third are installed in ships or submarines providing the energy for their locomotion; another third are used to generate electricity; the rest are used for research and the production of plutonium.

Over 15% of the world's electricity is currently produced by nuclear reactors; in certain countries this figure exceeds 50%.

The occupational and health hazards faced by workers in the nuclear generating industry are no greater than those met in other primary energy sectors, for instance, by underground coal miners.

Moreover, unlike the fossil-fuel industry, nuclear power produces no air pollution, acid-rain, oil-slicks, open-cast coal mines, leaking oil pipe-lines, rusty drilling rigs or abandoned refineries.

On the other hand, disposing of the radioactive waste that is unavoidably produced during the operation of nuclear reactors and decommissioning the reactors after the end of their useful lives, presents great problems.

These are probably technically soluble but the cost appears to be prohibitive.

The proliferation of the potential weapons grade fissile material produced in the reactors is another factor predisposing the wisdom of their use.

Since the 1980's, world-wide construction of new reactors for power generation has virtually ceased.

The reasons for this are many, doubts about the true cost of the electricity they produce and their safety being chief amongst them.

Notwithstanding these reservations, faced with (i) a growing world population and the demand for improved living standards; (ii) the inevitable future exhaustion of fossil fuels; (iii) the limits placed on the supply of agricultural energy by the availability of cultivable land and fresh water; and (iv) the practical and aesthetic restrictions on the wide-scale development of solar power, sometime in the near future there will probably be no choice but to return to nuclear processes for the required extra energy production.

The alternatives may be either the perpetuation of the present inequalities in world living standards or a degree of population control and social engineering that would undermine basic human freedoms.

The greatest threat to the safe operation of a power station is a sudden or undetected breakdown in the mechanism that extracts the heat from the reactor or furnace.

In a typical power station, nuclear or conventional, heat is being continuously generated at a rate of ~3000 megawatts.

This is sufficient heat to convert more than 1,000kg of water into steam each second.

Under normal operating conditions, this heat is extracted from the reactor or furnace by the coolant and transferred to the turbine-generators.

However, if the heat generated during just a few seconds was to accumulate in the reactor or furnace, it would either melt or explode.

Thus, an undetected or sudden breakdown in the system which circulates the coolant round the reactor or burners, or, alternatively, a loss by leakage or evaporation of the coolant itself, is the most dangerous event that can occur in a power station.

In reactors fueled by enriched uranium, such as the so-called Light Water Reactors-LWR of which the pressurized water reactor is an example, the coolant water is also the moderator.

Consequently, such reactors are inherently safe against a loss of coolant.

No action is required to close them down if coolant is lost.

This is because losing the coolant means losing the moderator, and without the moderator to slow the neutrons the only fission possible is fast-neutron fission, whose probability (nuclear cross-section) is some 500 times less than that of slow neutron fission.

Consequently, if it loses its coolant, an LWR reactor closes down of its own accord.

The reactor immediately becomes sub-critical and the chain reaction automatically ceases.

In the accident in the LWR at the Three Mile Island installation in 1979, a loss of coolant caused the almost instantaneous close-down of the nuclear chain-reaction.

Thus, there was never any danger of a run-away nuclear fission reaction.

However, a small proportion of the heat generated by a nuclear reactor (one or two percent) comes from the radioactive disintegration of the relatively long-lived fission fragments, the unstable medium-sized nuclei, that accumulate in its fuel elements.

These heat generating processes continue even after the reactor has closed down the fission chain reaction.

Consequently, unless appropriate measures are taken to continue cooling the reactor, even after the chain reaction has ceased, there is a danger that the assemblage will gradually heat up and ultimately even melt.

However, the probability of the melt self-organizing into a nuclear bomb is essentially zero.

When the cooling system in the reactor at Three Mile Island broke down, the containment structure enclosing the reactor prevented the spread of any significant amounts of radiation to the surroundings.

The reactor that exploded catastrophically at Chernobyl in April 1986 was designed for the concurrent production of electricity and weapons-grade plutonium.

It was constructed in such a way that its fuel elements could be removed and replaced after as short an interval as was required without closing down the reactor.

In order to achieve this capability, it was necessary to dispense with the containment structure.

Like all the early reactors designed to produce weapons grade plutonium, the moderator used in the Chernobyl reactor was pure graphite.

The coolant was water, which boiled as it flowed through pipes in the reactor. The steam thereby produced drove the turbine-generators.

Unlike a light water reactor, a Chernobyl-type reactor does not automatically close down in the event of a loss of coolant.

Just the opposite.

Hydrogen nuclei have an appreciable nuclear cross-section, $\sigma = 0.3$lb, for absorbing slow neutrons.

Consequently, the water coolant, H_2O, absorbs some of the neutrons in the reactor; water is a mild 'reactor poison'.

Hence, a loss of the water coolant results in an increase in the net neutron flux inside the reactor core and so enhances the chain reaction.

Thus, there is an imminent danger of an explosion or melt-down in the event of a sudden or undetected loss of the water coolant.

Its designers knew that reactors of the Chernobyl type are inherently unsafe against a loss of coolant.

Because of this, they incorporated an array of alarm systems as well as several automatic secondary (back-up) cooling and pumping systems into the plant.

Furthermore, in the event of the failure of these systems, there was a mechanism by which neutron absorbing cadmium rods would be automatically released into the reactor, shutting it down. In the search for still greater safety, emergency diesel generators had been installed in order to keep the coolant pumps working in the event of a loss of main power.

It had been further suggested that, in an extreme emergency, the angular kinetic energy of the power station's own rotating turbines could be used to generate the electricity needed to keep the coolant pumps working for the few minutes until the emergency generators came on line.

The catastrophe occurred during an experiment that was being performed to investigate this possibility.

It was originally intended to carry out the experiment on the spring day the reactor was to be shut-down for its annual maintenance.

The experiment was to have been conducted with the reactor operating at about 30% (700-l000MW) of its full capacity (~3000MW), a level at which it was still controllable.

However, because of a sudden bout of cold weather, at the last moment, the shut-down was put off for a few days.

It was re-authorized when the weather improved, but because of the pressure of time, the reduction from full capacity was then carried out too quickly.

As a result, a quantity of 'poisonous' Xe-135 accumulated in the reactor.

At this stage the reactor was operating at just 1% (~30MW) of its capacity.

Instead of immediately aborting the experiment, the operators attempted to 'rekindle' the reactor by withdrawing some of its cadmium control rods.

After about two hours, they had managed to coax it back up to about 7% (200MW) of its maximum capacity.

Because of the Xe-135 that had accumulated in the reactor, this was the highest they could achieve.

At this point the experiment was started.

All the alarm systems and the many automatic safety devices were first neutralized.

Leaving them on would have closed down the reactor automatically the moment the main power supply was cut off.

In accordance with the original plan, the flow of water coolant into the reactor was then increased.

Since water is a reactor poison, this had the effect of slowing the chain reaction.

To compensate for this and in an attempt to maintain power, most of the remaining cadmium control rods were now withdrawn.

This left the neutron flux dependent on the liquid water coolant in the reactor alone.

Any loss of coolant would release so many more neutrons for the chain reaction, that the reactor's multiplication factor would almost instantly rise above the permitted level, putting it beyond control.

At 01:22 local time on April 26, the flow of coolant into the reactor was reduced but no control rods were re-inserted.

Ignoring the alerts to shut down the reactor at once that were now flashing on their computer terminals, the operators continued with the experiment.

At 01:23:04, the supply of steam was cut off from the turbines, leaving them to rotate freely.

It was in this state that their kinetic energy was to have been utilized to operate the coolant pumps.

But by now, the chain reaction in the reactor was going out of control.

Reducing the coolant flow without re-inserting control rods was catastrophic.

With less 'poison' flowing through the reactor, the chain reaction could speed up.

This in tum boiled off more of the water, increasing the steam pressure inside the reactor.

But with its flow reduced, there was less water to be boiled and to cool the reactor.

But less water meant less 'poison' and so the chain reaction could proceed even faster.

This vicious circle of cause and effect was driving the reactor out of control.

The control rods were now re-inserted, but to no avail; the heat production in the reactor continued to grow.

At 01:23:40 the order was given to release the emergency rods into the reactor, but they jammed.

The shafts down which they were supposed to drop had been bent out of shape by the heat generated in the reactor.

At about 01:24 two explosions were heard and fires broke out throughout the reactor core.

It is estimated that at the moment of these explosions parts of the reactor were operating at a capacity 20 to 100 times its normal maximum rating.

Within two hours, the fires in the 1500 tons of graphite that had been the reactor's moderator were extinguished.

During the following days, some 5000 tons of boron compounds (to absorb the neutrons and ensure that there was no danger of the mass becoming supercritical), sand, stone and lead were dropped from helicopters onto the smoldering debris.

Most of the pilots who carried out these sorties, as well as a number of the firemen, received fatal doses of radiation and have since died.

By May 6 1986 the level of radiation from the site had been reduced to reasonable levels.

Since then, the reactor site, together with the debris from the explosion, has been entombed in a concrete sarcophagus.

What happened at Chernobyl was not an accident.

It was the result of poor design and an inflexible management philosophy.

Instead of relying on the natural imperatives of physics to ensure the plant's safety, its designers put their trust in planning.

This was the underlying philosophy of their political system at the time.

But whereas Nature, though unforgiving, can never go wrong, planning is always subject to Murphy's Law: "If it can go wrong, it will!"

Within five years of the events at Chernobyl, the citizens of that unfortunate town, together with their fellow country-men, had overturned the system.

Part 6 Selected Applications

Part 6.1 The Laser

The word 'Laser' is an acronym of Light Amplification by the Stimulated Emission of Radiation.

Since their invention in 1960, lasers of various types have found numerous applications in research, enterprise and technology.

The key features of the light produced by a laser are: its almost pure monochromacity; its coherence; its high intensity and, fourthly, its low beam divergence.

For example, the beam from a typical l0mW helium-neon laser has a monochromacity of 632.81±0.02nm, a beam diameter at source of only 0.5mm and a beam divergence of just 0.00lradians; at a distance of 100 meters from its source the beam's diameter is just 10cm.

To gain an understanding of laser action, we must first review the mechanisms by which atoms and molecules absorb and emit radiation.

6.1.1 The Spontaneous and Stimulated Emission of Radiation

Excited atoms and molecules can emit radiation (photons) in two different ways: spontaneously or as a result of being induced to do so.

The spontaneous emission of radiation from an excited atom or molecule is a pure quantum effect that has no classical analogue. Ideal classical oscillators do not spontaneously lose energy; an undamped pendulum would go on swinging for ever.

Spontaneous emission: $\text{Atom}^* \rightarrow \text{Atom} + \text{photon}$

Spontaneous emission is a chance phenomenon; the emissions occur haphazardly, at unrelated times. Consequently, spontaneously emitted radiation is incoherent.

Or, to put it another way, spontaneously emitted photons have random phases.

Excited atoms and molecules can also be induced to emit radiation by the impact of radiation (photons) upon them, radiation from another source which strikes the excited atom or molecule and triggers the ejection of a photon from it.

This second mechanism is called *stimulated emission*.

Stimulated emission: $\text{Atom}^* + \text{photon} \rightarrow \text{Atom} + 2\ \text{photons}$

Stimulated emission only occurs if the energy of the incident photon equals one of the transition energies in the excited atom or molecule, i.e., if the frequency of the incident photon matches that given by Bohr's equation, $f_{ij} = (E_i - E_f)/h$, where E_i and E_f are the energy levels of two stationary states.

The new photon emitted by the stimulated emission of an excited atom or molecule is always identical to the triggering photon; it moves in the same direction and has the same energy, polarization and phase as the incident photon (Fig 6.1).

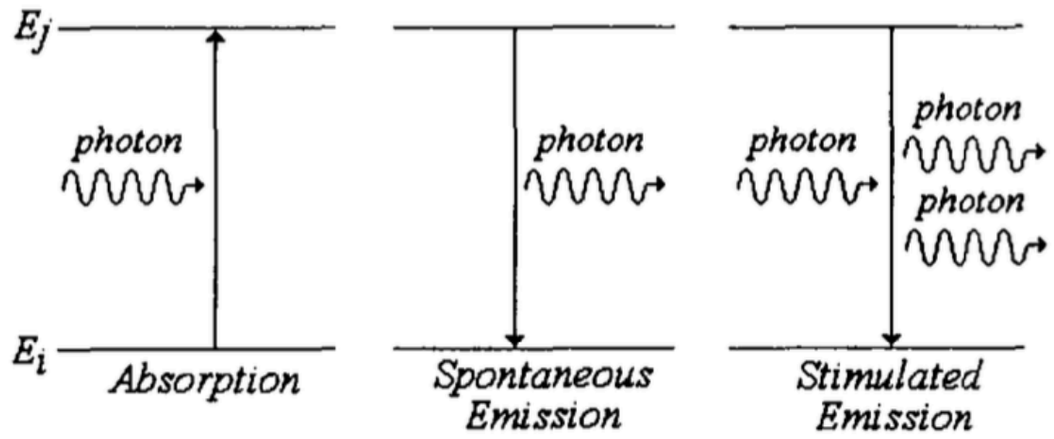


Fig 6.1 Energy level diagrams for the processes of absorption, spontaneous emission and stimulated emission.

The same essential event, viz. the collision of a photon with an atom or molecule, is involved in the stimulated emission of radiation as in the absorption of radiation.

They also share a classical analogue: the phenomenon of resonance.

In both cases, we say that there is resonance between the radiation and the atoms or molecules.

In classical mechanics, when a periodic force and an ideal oscillator both possess the same frequency, they are said to be in resonance.

If the periodic force is coupled to the oscillator, they will exchange energy.

Whether the oscillator gains or loses energy as a result, depends on the phase difference between the periodic force and the oscillator's vibrations.

If the two are in phase, the oscillator gains (absorbs) energy; if the phase difference between them equals *n*, the oscillator loses (emits) energy.

Thus, absorption and stimulated emission differ essentially only in terms of phase.

A gain or loss in the energy of an atomic or molecular oscillator is manifested, respectively, as the absorption or stimulated emission of radiation.

The former corresponds to a 'jump' and the latter to a 'fall' between the energy levels.

That excited atoms and molecules emit radiation by stimulated as well as spontaneous emission can be shown by considering the dynamic equilibrium that exists in cavity radiation.

Suppose, for simplicity's sake, that the system comprises just two energy levels, E_1 and E_2, that are occupied by N_1 and N_2 particles, respectively.

Assuming that the distribution of the particles between the two levels follows the Maxwell-Boltzmann rule, the ratio between their populations will be:

$$N_1/N_2 = e^{(E_2-E_1)/kT} = e^{hf/kT} \tag{6.1}$$

where *f* is the frequency of the photons absorbed or emitted in transitions between the two levels.

At equilibrium, the ratio N_1/N_2 must on average be constant, i.e., the rates of the upward and downward transitions must be equal.

The rate at which upward transitions occur clearly depends on the intensity of the radiation in the cavity: the greater the energy density of the radiation, the more collisions there will be between photons and atoms in the cavity's walls and the greater the number of upward transitions that occur.

On the other hand, spontaneous emission is unaffected by changes in the energy density of the radiation in the cavity; there is no way it can adjust to changes in this density.

Hence, spontaneous emission cannot of itself balance these upward transitions.

It follows, that in order for the upward and downward transition rates to balance, an emission mechanism that depends on the energy density must also exist.

This mechanism is the stimulated emission that occurs when photons strike excited atoms in the cavity's walls.

It can be shown that, at equilibrium, the transition probabilities, P_{St} and P_{Ab}, of stimulated emission and absorption, respectively, are equal, i.e., $P_{St} = P_{Ab}$

This result was first obtained by Einstein in 1917.

He also showed, that the ratio, P_{Sp}/P_{St} between the transition probability for spontaneous emission, P_{Sp}, and that for stimulated emission, P_{St}, is given by:

$$P_{Sp}/P_{St} = \kappa_f\left(e^{hf/kT} - 1\right) \tag{6.2}$$

where f is the frequency the emitted radiation and κ_f is the average photon density in the system.

Thus, at high frequencies, where hf >> kT, spontaneous emission is the more probable of the two emission mechanisms; this is the case with electronic transitions in atoms and molecules as well as for nuclear radiative transitions.

However, at low frequencies, where hf << kT, stimulated emission becomes more important; this is the case in microwave transitions and nuclear spin transitions.

6.1.2 Laser Action

The two coherent photons that emerge from a stimulated emission can go on to induce further emissions in other excited atoms in the system.

Thus, in principle, the conditions exist for propagating a chain reaction that can amplify the intensity of the radiation.

However, as we have seen, the probability that a photon will induce the emission of a second photon from an excited atom is equal to the probability that it will itself be absorbed by an atom in the lower state.

Thus, a chain reaction of stimulated emission can only develop if the number of atoms in the excited state is greater than the number in the lower state.

Achieving such a population inversion is the first requirement in the construction of a laser device.

Because of the equal probabilities of absorption and stimulated emission, a population inversion cannot be achieved in systems that comprise only two energy levels; the system must have at least three energy levels.

The simplest configuration that can produce laser action is shown schematically in Fig 6.2.

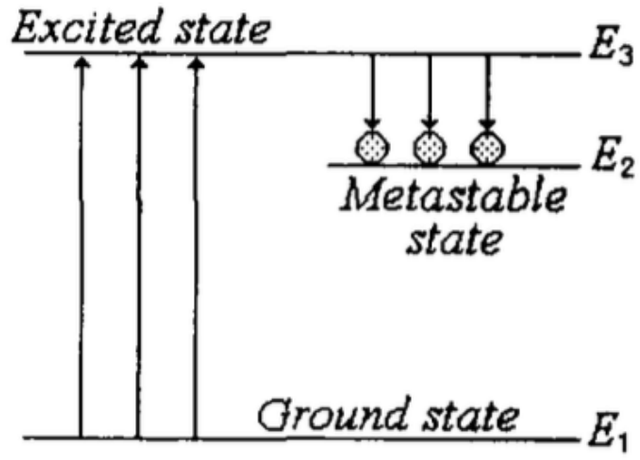


Fig 6.2 Population inversion in a three level system. As a result of the accumulation of atoms in the metastable state its population exceeds that of the ground state.

Atoms, initially in the ground state, E_1, are excited into the E_3 state from which they quickly ($\sim 10^{-9}$s) fall by non-radiative transitions into the metastable state, E_2.

The E_2 state is typically a triplet state (**S** = 1) that has an average life-time of about a millisecond; the probability of spontaneous emission from this state is low.

Thus, atoms rapidly accumulate in the metastable E_2 state and within a short time its population overtakes that of the ground state, E_1.

Having achieved a population inversion, the stimulated emission chain reaction will only propagate if each of the photons emitted by the atoms in the metastable state triggers the emission of at least one more photon before escaping from the assembly.

This is the second requirement for the construction of a laser device.

It can be satisfied:

1. By making the laser component in which the population inversion is achieved long enough to ensure that each photon strikes an atom in the metastable state before escaping (most X-ray lasers work this way).
2. By installing a pair of parallel mirrors at opposite ends of the component and reflecting the photons emitted parallel to the axis of the device to and fro along its length (Fig 6.3). In essence the component is repeatedly folded up between the mirrors. The chain reaction propagates with each passage of the photons between the mirrors as more and more atoms are triggered into emitting a photon and the photon flux in the component increases.

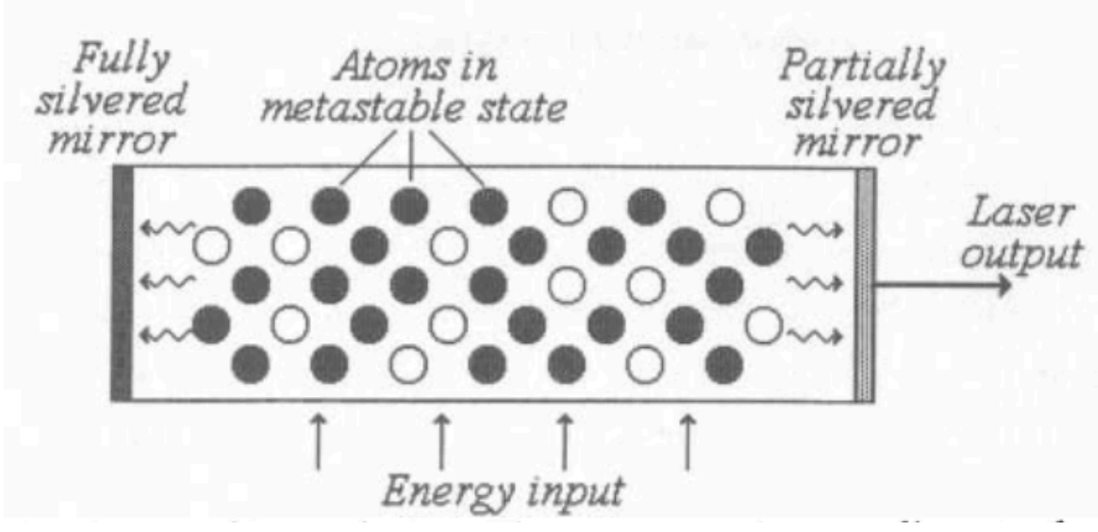


Fig 6.3 A schema of laser design. The atoms are 'pumped' up to the metastable state by the energy input. The parallel end mirrors reflect the stimulating photons to and fro through the inversed population of atoms. The laser output emerges from the partially silvered end as a narrow beam.

If the distance between the mirrors is equal to an integer number of half wavelengths of the light emitted by the atoms, a highly monochromatic standing wave, whose amplitude grows as more and more photons accumulate, is generated along the component's length.

The component may be regarded as an optical resonator.

If the mirror at one end is only partially silvered, light ultimately leaks out from it as a narrow highly monochromatic coherent beam.

6.1.3 The Ruby Laser

Laser action with visible light was first achieved by Maiman in 1960 using a cylindrical ruby whose ends had been polished and silvered.

Ruby is a crystal of aluminum oxide (Al_2O_3) doped with a small amount of chromium.

The chromium atoms in the crystal, present in a concentration of about 0.1%, possess a three level system of the type shown in Fig 6.2.

The energy difference, E_3 - E_1, of the chromium system corresponds to a photon wavelength of 550nm (Fig 6.4).

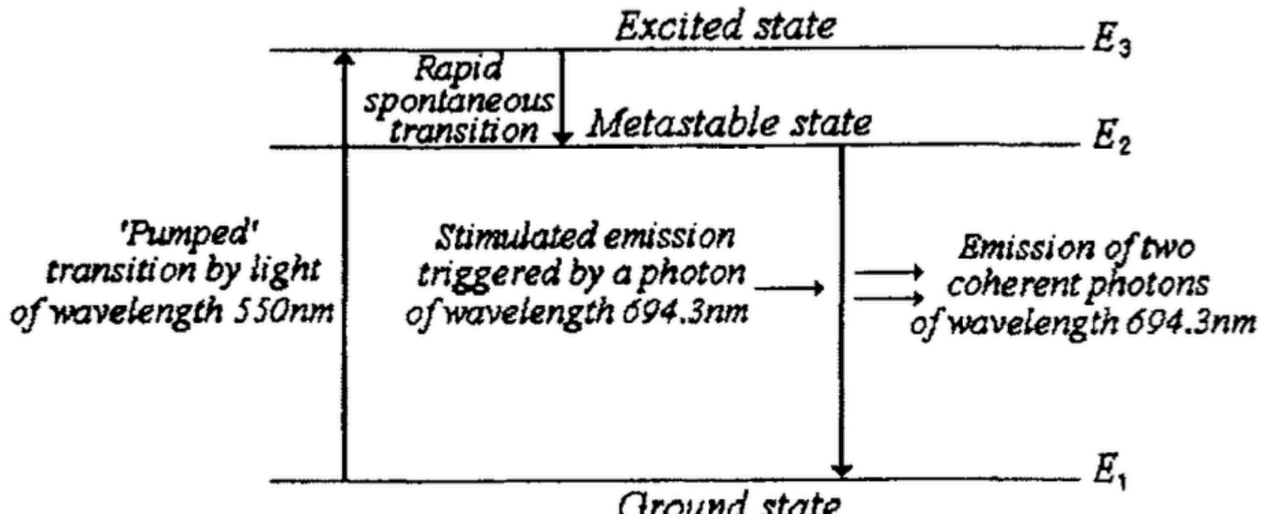


Fig 6.4 Diagram of the transitions that occur in a ruby laser. The chromium atoms are raised to the higher excited state by absorbing a 550nm photon. They rapidly and spontaneously fall into the metastable state from which they return to the ground state by stimulated emission.

Maiman obtained the required population inversion of the chromium atoms by 'pumping' the ruby crystal with flashes of green/yellow light of wavelength 550nm.

The metastable E_2 state has a lifetime of about 3 milliseconds after which the excited atom returns to its ground state with the emission of a 694.3nm (red) photon.

The stimulated emission chain reaction in the ruby crystal is initiated by the few photons emitted spontaneously from chromium atoms in the metastable state.

In order to propagate the chain reaction, these 694.3nm (1.79eV) photons are reflected to and fro between the polished and silvered ends of the ruby crystal.

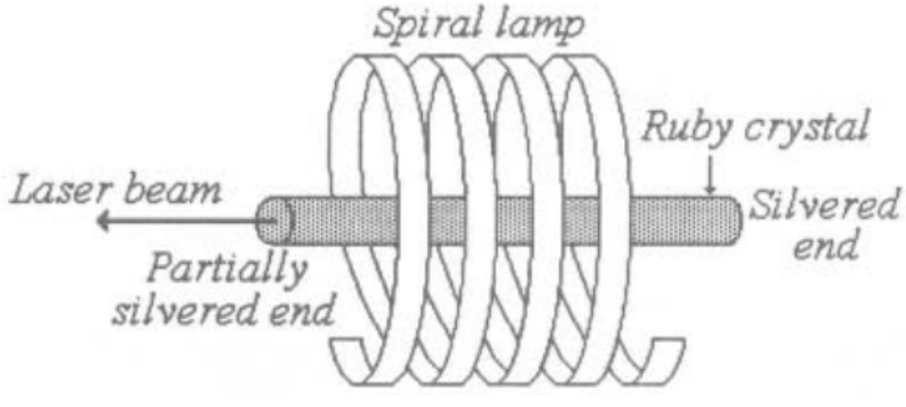
The chain reaction propagates with each passage of the photons as more and more chromium atoms are triggered into emitting a photon.

The amplitude of the standing wave set up in the crystal grows as the photons accumulate.

Ultimately light leaks out from the partially silvered end as a narrow highly monochromatic coherent8 beam.

‘Typically, about 1% of the photons incident on the partially silvered surface emerge from it (Fig 6.5).

Fig 6.5 Ruby laser. Intense flashes of light from the spiral lamp excite the chromium atoms in the cylindrical crystal into a metastable state; a population inversion is achieved. The repeated reflections of the emitted light from the silvered ends of the crystal amplify the light intensity in the crystal until it emerges as a pulse of monochromatic coherent light from the partially silvered end of the crystal.



The ruby laser is an example of a solid state laser.

Each pumping flash of light lasts for a few milliseconds during which the device emits a number of rapid pulses of laser light each of about 0.2 - 0.3ms duration.

The cross-sectional area of its beam is about 0.1mm^2 and so at a mean operating power of 2kW the intensity of each pulse is about $2x10^{10}w/m^2$.

Even greater intensities can be obtained by suppressing the stimulated emission until most of the 'pumping' energy of the light flash has been absorbed by the system.

This can be achieved by replacing one of the fixed mirrors at the crystal's ends by a vibrating or rotating mirror that is phased with the pumping lamp to produce the resonant conditions in the ruby only when a highly inverted population has been established.

In this way single pulses lasting just l00ns in which all the energy absorbed during a single pumping flash is emitted can be produced.

Such ultra-short pulses have a power of ~ 10MW.

The ruby laser has a number of drawbacks:

* **Because of the heat generated in the crystal during its operation, it cannot be operated continuously. It must be pumped by spaced flashes of light.**
* **For laser action to occur, at least half of the chromium atoms in the crystal must be in the metastable state. This is the reason why the ruby laser emits light in rapid pulses during excitation and not continuously. Once the chain reaction has started, the metastable state depopulates faster than it is repopulated. As a result, after each pulse of laser action several tenths of a millisecond are needed to restore the population of the metastable state.**

Both of these disadvantages are avoided in the four level helium-neon gas laser.

6.1.4 The Helium-Neon Laser

In the helium-neon laser, the crystal is replaced by a discharge tube containing a mixture of helium (85%) and neon (15%) at low pressure.

The tube has parallel mirrors at its ends just like the crystal in the ruby laser.

The laser is pumped by the electrical discharge across the tube.

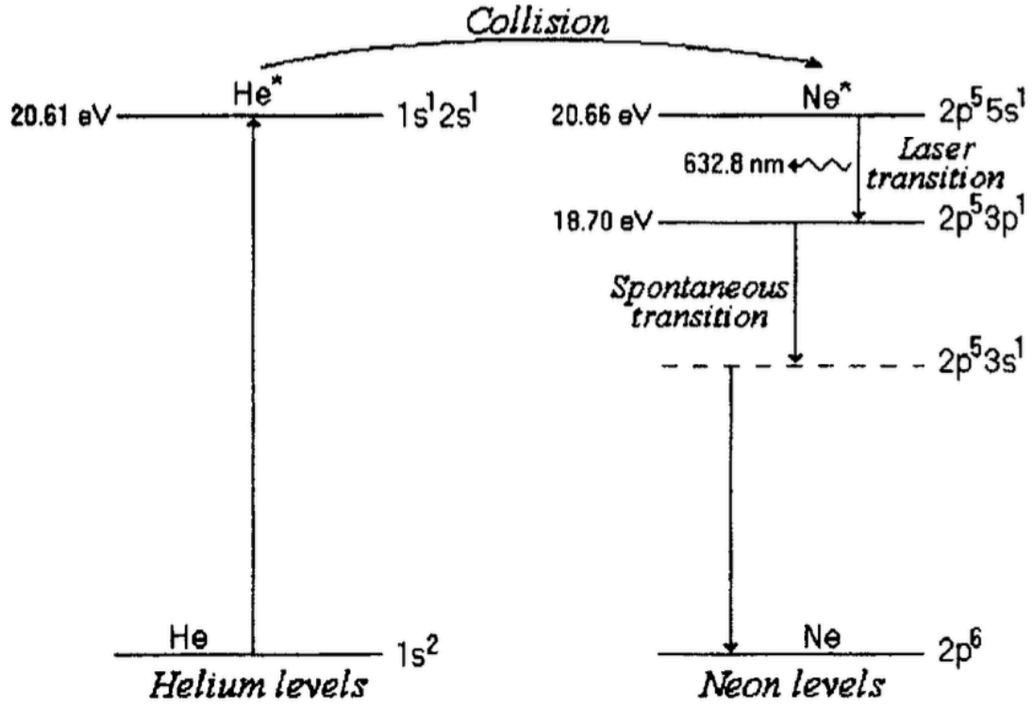


Fig 6.6 The helium-neon laser. Energy is continuously supplied to the neon atoms by the more abundant excited helium atoms. This maintains the population of the $2p^5 5s^1$ state. The population of the $2p^5 3p^1$ state is always negligible due to its short life.

The population inversion in a helium-neon laser does not involve the ground state and so its operation does not require that at least 50% of the atoms involved in the laser action be in the metastable state.

The inversion is achieved by stages.

Initially, the more abundant helium atoms in the discharge tube are electrically excited from their ground state up to various higher states from which they spontaneously and rapidly fall into the metastable $1s^1 2s^1$ state of energy 20.6leV.

Neon atoms possess an excited state, $2p^5 5s^1$, of almost the same energy, viz. 20.66eV.

In collisions between helium atoms in the excited metastable state and neon atoms in the ground state, the helium atoms transfer their energy to the neon atoms.

In the process, the neon atoms are raised to the $2p^5 5s^1$ state and the helium atoms return to their ground state.

$$\mathbf{He^* + Ne \rightarrow He + Ne^*}$$

Neon atoms in the $2p^5 5s^1$ state can fall by stimulated emission to the short-lived $2p^5 3p^1$ state of energy, 18.7eV, from which they spontaneously and rapidly drop to the $2p^5 3s^1$ state.

The neon atoms finally return to their ground state from the $2p^5 3s^1$ level.

Because of the continuous transfer of energy from the excited helium atoms to the neon atoms and because of the short life of the $2p^5 3p^1$ state, the $2p^5 5s^1$ state will always be more highly populated by the neon atoms than the lower $2p^5 3p^1$ state.

This is the population inversion that produces the laser action (Fig 6.6).

The visible light emitted by a helium-neon laser has a wavelength of 632.8nm corresponding to a photon energy of 20.66-18.7 = 196eV.

The output of a typical helium-neon laser is about 5 - l0mW; they are low power lasers.

Other gas lasers can produce continuous beams of very high intensities; powerful carbon dioxide lasers (~200W) can easily cut through sheet metal.

Short pulses (~50ns) with a peak power of 50MW can be produced by chemical lasers in which the population inversion is achieved by exciting the gas molecules participating in an energetic chemical reaction into the metastable state.

6.1.5 Laser Applications

The different uses to which laser light is put can be grouped under the headings of its key characteristics: monochromacity, coherence, high intensity and low beam divergence.

a. <u>Monochromacity</u>

The lines in atomic emission spectra are not absolutely monochromatic; if they were, the lines would be infinitely narrow and invisible.

Typical spectral lines have a variance of about ±0.001%.

A red spectral line might comprise wavelengths in the range 700±0.007nm; its linewidth is of the order 10^{-2}nm.

However, laser light can be produced with a linewidth as fine as 10^{-13}nm.

Stabilized lasers that maintain a constant wavelength to within 2 parts in 10^{12} have enabled the measurement of the speed of light and other physical constants with unprecedented accuracy.

The availability of almost pure monochromatic laser light has had a great impact on spectroscopy, particularly since the invention of the tunable dye laser.

In this device the length of the optical resonator in which the population inversion is achieved can be finely adjusted.

The active medium contained in the resonator is an organic fluorescent pigment.

The molecules of organic pigments have large numbers of closely spaced energy levels and so, in a given range, they can emit light over an almost continuous spectrum of wavelengths.

Excitation of the fluorescent pigment into its metastable states and adjustment of the length of the optical resonator, enables selection of monochromatic laser light of any wavelength within the given range.

Scanning a substance with the range of frequencies available from a tunable laser, reveals the precise frequencies at which it absorbs; at these frequencies the attenuation of the laser beam passing through the sample is greatest.

Tunable lasers have also been used for isotope separation by means of selective ionization.

Notwithstanding the very small, almost negligible, differences between their energy levels, the frequency of the laser beam can be so finely tuned that it will excite one isotope in a sample but not the others.

b. <u>Coherence</u>

Ordinary light sources, such as white hot objects (the filaments in electric bulbs) or discharge tubes, emit light in short wave trains, each about a meter long or ~10^{-8}s.

Lasers, however, can emit wave trains hundreds of kilometers in length (~10^{-3}s).

For this reason and because of the great intensity of their light, lasers are an excellent source for experiments on the interference and diffraction of light.

The wave trains of lasers are so long that it has even been possible to observe the phenomenon of beats between two independent laser sources, a phenomenon well known in sound but unobservable in light before the advent of the laser.

The use of the holographic technique of imaging, invented by Gabor in 1948, has been greatly extended by the availability of coherent high intensity laser light.

Traditional photographic imaging has the disadvantage that the camera lens must be focused on a particular plane in the field of view.

Only this plane is in focus in the photograph taken; all other planes are out of focus to a greater or lesser extent. Holography overcomes this limitation.

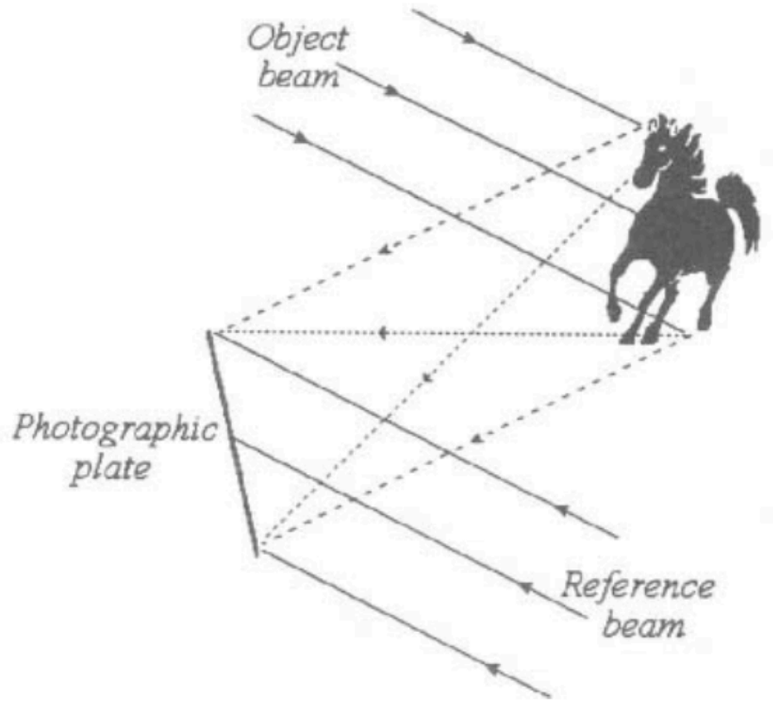
The essential ideas underlying holography are:

1. **that an interference or diffraction pattern contains all the information needed to construct an image of the physical system in which it was formed;**
2. **that this information can be stored for future retrieval as a photographic etching of the interference or diffraction pattern called a *hologram*.**

To prepare a hologram, a laser beam is first split into two beams by means of a suitable optical arrangement of half-silvered mirrors.

One of the beams, called the 'reference beam' is then aimed directly at a high resolution photographic plate, whereas the second beam, called the object beam', is aimed at the object whose image is to be recorded (Fig 6.7).

Fig 6.7 The laser light scattered by the object (the horse) interferes on the photographic plate with the light in the reference beam.



The light scattered by the object strikes the photographic plate where it combines with the light from the reference beam to produce an interference pattern (rather than the usual areas of light and dark intensity that appear on a conventional photographic negative).

In the absence of the object the two beams would combine on the plate to give the typical two-slit interference pattern associated with Young's experiment.

The presence of the object modulates this basic pattern just as a signal modulates the carrier wave in radio or telecommunications.

In order that just one interference pattern be formed on the plate, the highly coherent light of a laser must be used in preparing the hologram.

Using non-coherent light would result in 'noise' instead of a clear signal.

After developing, the photographic plate acts as a diffraction grating.

An image of the original object can be re-created by passing a beam of coherent light, generally but not necessarily laser light from the original laser, through the grating.

Two strongly diffracted beams are obtained: one can be projected onto a screen to give a real image of the object; the other appears as a virtual stereoscopic image when looking through the transparent photographic plate (Fig 6.8).

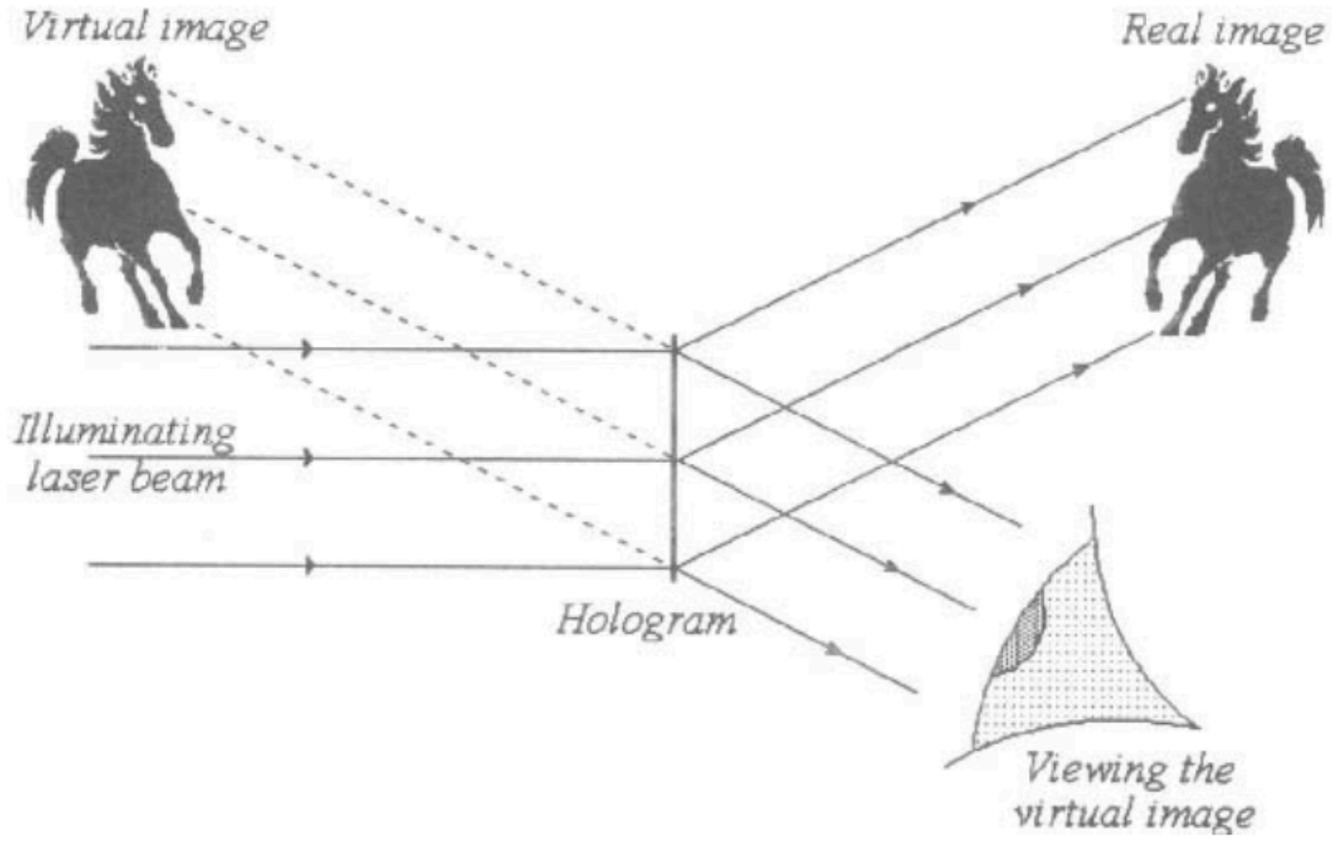


Fig 6.8 The hologram transparency is illuminated by a beam of laser light. The first order diffracted beam provides a real image of the object on one side of the zero and a high quality virtual image on the other.

The Because there is no correspondence between points on the object and points on the hologram, all parts of the image can be seen by viewing through any part of the hologram.

Thus, all parts of the image can be obtained from just a fragment of the hologram.

Full color holograms, that can be viewed in ordinary incoherent white light, can also be produced.

For example, those embossed on credit cards to prevent their forgery, it being virtually impossible to exactly copy them.

The full color holograms are etched on a thick photographic emulsion using three laser beams, each in one of the primary colors - red, green and blue.

c. <u>High intensity and low beam divergence</u>

The unfocused beam from a low intensity l mW helium-neon laser has a brightness equal to that of sunlight (~l kW/m^2).

However, laser beams can in principle be focused to a width no greater than their wavelength with a corresponding increase in intensity .

Thus, the beams from I W lasers can be focused to an intensity of 10^9 kW/m^2 and pulsed ruby lasers have produced beam intensities as high as 10^{17} kW/m^2; beams of this intensity can ionize the air and puncture metal sheets.

These high intensities have found many applications in industry, medicine and research.

In industry, focused laser beams have been used to cut metals and to weld metal parts together, for the accurate computer controlled cutting of cloth and for the spot hardening of sensitive mechanical components.

The use of a focused laser beam instead of a scalpel in surgery has the advantages of total sterility and the instantaneous cauterizing of severed blood vessels, thus reducing bleeding.

Detached retinas are routinely 'welded' back into place using lasers.

Using bundles of optical fibers (endoscopes), laser light can be introduced through the natural orifices of the body, conducted around its inner organs and aimed at specific locations within the body without the need for conventional surgery.

The optical fibers comprise long thin cylindrical crystals of a transparent material such as glass.

Light introduced at one end of the fibre is repeatedly reflected off its inside surface as it propagates along its length.

Thus, the light is kept trapped inside the fibre until it emerges at the other end with little attenuation.

Because of their low beam divergence, laser beams can be used to determine the location or movement of distant objects.

For example, lasers have been used to monitor the movement of dams as the level of the water in their reservoir rises and falls as well as the geological movements associated with plate tectonics.

By accurately measuring the time taken for a pulse of intense laser light reflected from mirrors positioned on the moon by the Apollo astronauts to return to earth, the distance from the earth to the moon has been measured to an accuracy of ±10cm.

The most widespread use of laser light is in the field of information processing.

In all these applications, the information is encoded as a modulation of the laser light.

A familiar application is the optical scanner, of the type seen at supermarket check-outs, that reads bar codes.

The laser beam scans the bar code and is reflected from it.

The intensity of the light reflected from the bar code varies in time with the pattern of its white and black lines.

These variations are detected and converted into an electrical signal, which is then fed into a computer.

Low power helium-neon lasers are typically used in these scanners.

Information is stored on the compact optical discs used as read-only memories for computers and to provide digital audio and video signals, as a series of small pits or holes in the disc's reflecting surface.

As the disc spins, the light from a small semiconductor laser is focused onto its surface.

The pattern of pits is read by detecting the modulations in the light reflected from the disc's surface and converting them electronically into a digital signal.

Since the laser beam can be made as narrow as a single wavelength, the pits on the disc need be no more that about one micron in size.

In this way, vast amounts of information can be stored on a single disc.

Information is stored on the compact optical discs used as read-only memories for computers and to provide digital audio and video signals, as a series of small pits or holes in the disc's reflecting surface.

As the disc spins, the light from a small semiconductor laser is focused onto its surface.

The pattern of pits is read by detecting the modulations in the light reflected from the disc's surface and converting them electronically into a digital signal.

Since the laser beam can be made as narrow as a single wavelength, the pits on the disc need be no more that about one micron in size.

In this way, vast amounts of information can be stored on a single disc.

Part 6.2 The Mossbauer Effect

One of the most exciting aspects of scientific research is the way that technical advances made in one field can impact on other seemingly unrelated fields.

The life sciences were twice revolutionized by advances in microscopy.

First, by the invention of the optical microscope which revealed the existence of previously unseen forms of life; **second,** by the invention of the electron microscope which revealed the existence of the viruses and of cellular architecture.

Experimental evidence for Einstein's theory of general relativity and for the existence of the exotic bodies such as 'black-holes' whose existence it posits, has been hard to come by.

The extreme accelerations and strong gravitational fields under which its effects are clearly manifested only exist in deep space or in deep time.

The detection of its predicted effects on earth requires the ability to make measurements of gravitational effects at an accuracy greater than one part in 10^{14}.

This only became possible with the discovery by Mossbauer in 1958 of a technique that enabled such measurements to be made.

The technique, known as the *Mossbauer Effect*, is associated with the emission and absorption of γ-ray photons.

As part of their decay process, many radioactive nuclei emit high energy electromagnetic radiation - γ rays.

These γ rays are emitted as the nuclei fall from an excited nuclear state to a lower one.

However, under ordinary conditions, nuclei already in the lower state do not absorb the γ rays emitted by identical excited nuclei as they drop into that state, i.e., though the nuclei are identical, there is no resonance between them.

This contrasts with the situation in atoms and molecules.

Generally, atoms and molecules in the ground state do absorb the radiation emitted by identical excited atoms and molecules as they drop into that state, i.e., there is resonance between identical atoms and molecules.

Elucidating the conditions under which resonance does or does not exist between identical nuclei, atoms and molecules, requires a prior understanding of both the nature of spectral lines and the mechanics of photon emission and absorption.

6.2.1 The Width of Spectral Lines

Spectral lines are not absolutely monochromatic; they are not composed of just the single photon frequency, $f_{if} = (E_i - E_f)/h$, given by Bohr's equation.

If they were they would be infinitely thin and, therefore, invisible.

What then are the reasons for the finite width (broadening) of spectral lines?

A primary cause of the broadening of spectral lines is the thermal motion of the particles emitting or absorbing the electromagnetic radiation.

For example, as a result of this random motion, some of the photons are emitted from particles moving towards the observer and others from particles moving away from him.

This, in turn, affects the frequencies detected by the observer due to a Doppler effect.

Some of the photons appear to have a frequency slightly greater than f_{if} and some a frequency slightly less, depending on the direction and magnitude of the velocity of the emitting particles.

The result is a broadening of each spectral line. The effect is particularly pronounced in gas spectra.

However, even after allowance is made for such effects, spectral lines are still found to have an irreducible finite width.

They always extend over a range of frequencies greater than the single frequency predicted by Bohr's equation for the particular transition.

This inherent broadening of spectral lines is explained by Heisenberg's uncertainty principle, $\Delta E\, \Delta t \approx \hbar$, according to which there is no way of knowing exactly how long a particle will remain in an excited state before it emits the photon.

All that can be determined is the average lifetime of the state, within the limits of an uncertainty, Δt.

For the same reason, the energies of stationary states (their energy levels) are not precisely known; there is always an inherent uncertainty, $\Delta E \approx \hbar/\Delta t$, called the energy width, associated with them -see Fig. 6.9.

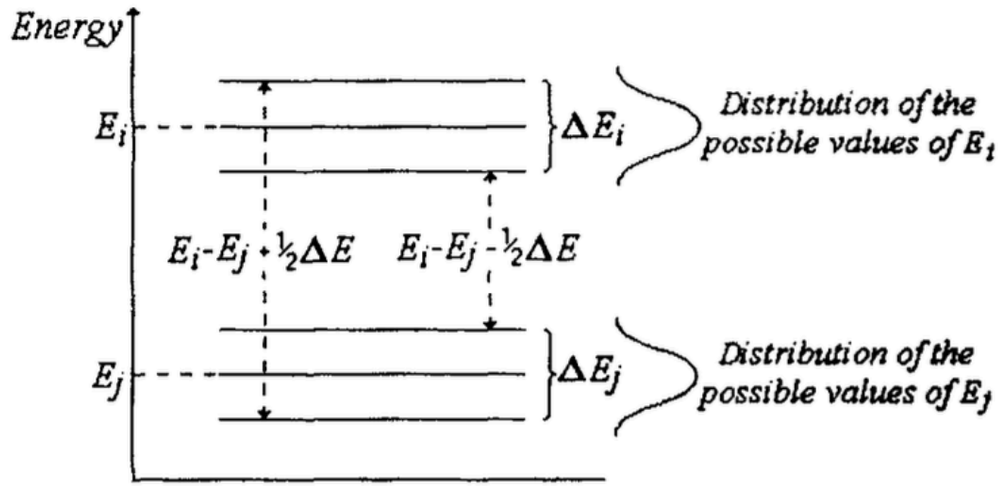


Fig.6.9 The variance and energy widths, ΔE_i and ΔE_j of the stationary states i and j, respectively. ΔE_i is the uncertainty in the value of E_i; and ΔE_j is the uncertainty in the value of E_j : $\Delta E = \Delta E_i + \Delta E_j$.

It can be reasonably assumed, that the uncertainty, Δt, in the average lifetime of a stationary state is of the order of magnitude of the state's average lifetime.

It follows, that the energy width of a state will be inversely proportional to its average lifetime, i.e., the less stable a state, the broader its energy width.

Hence, whereas unstable excited states have finite energy widths, the energy width of the ground state (the most stable state) can be taken as zero.

Consider two stationary states, i and j. Because of the variance (uncertainty) in their energy levels, the difference between their energies does not have a single precise value.

On the contrary, it can have any value in the range

$$\left(E_i - E_j\right) \pm \Delta E/2 \tag{6.3}$$

where $\Delta E = \Delta E_i + \Delta E_j$ is the sum of the energy widths.

The photons emitted or absorbed in transitions between these states will exhibit a similar range of energies.

Hence, even in the absence of any Doppler broadening, resonance is not limited to the unique frequency, f_{ij}, given by Bohr's equation.

All the photons whose frequencies fall in the range

$$f_{ij} \pm {}^{\Delta E}/_{2h} \tag{6.4}$$

where f_{ij} is the transition frequency given by Bohr's formula, may be absorbed or emitted.

This results in an inherent or 'natural' broadening of spectral lines.

From the fundamental uncertainty expression, $\Delta E\, \Delta t \approx \hbar$ and the relationship $\Delta E = h\Delta f$ the *natural width*, Δf, of a spectral line is given by

$$\Delta f \approx {}^{1}/_{(2\pi\Delta t)} \tag{6.5}$$

6.2.2 The Mechanics of Photon Emission and Absorption

We now turn to the mechanics of a radiative transition between the two stationary states, i and j.

Consider a particle of mass m, initially at rest in state i, that emits a photon.

The particle recoils backwards as it drops to the lower j stationary state.

Since the particle was initially at rest, its total linear momentum before the emission of the photon equalled zero.

The conservation of linear momentum requires that after the emission of the photon the sum of its linear momentum, $p_{particle}$, and that of the photon, p_{photon}, must also be zero:

$$p_{particle} + p_{photon} = 0 \tag{6.6}$$

Thus, the particle and the photon emitted from it move apart in opposite directions, each possessing an equal amount of linear momentum.

The linear momentum of a photon is given by $p_{photon} = E_{photon}/c$.

Since, after the emission, the photon and particle possess equal amounts of linear momentum, this must also be the magnitude of the linear momentum of the particle.

Hence we have:

$$|p_{particle}| = |p_{photon}| = E_{photon}/c \tag{6.7}$$

The total energy of the system after the emission of the photon is the sum of three quantities:

(i) the particle's energy in its new stationary state, E_j;

(ii) the kinetic energy of recoil of the particle, K;

(iii) the energy of the emitted photon, E_{photon}.

Energy conservation requires that the sum of these three energies equal the energy, E_i, possessed by the particle in its original stationary state, i.e.,

$$E_i = E_j + K + E_{photon} \tag{6.8}$$

Thus, the energy of the photon emitted when the particle falls from the i state to the j state is:

$$E_{photon} = \left(E_i - E_j\right) - K \tag{6.9}$$

It can be shown, by similar reasoning, that when a photon strikes the particle and is absorbed, and as a result the particle rises from the j state to the i state, the energy of the absorbed photon is:

$$E_{photon} = \left(E_i - E_j\right) + K \tag{6.10}$$

Hence, the energy of the photons absorbed when the particle rises from state j to i is greater by an amount equal to 2K than that of the photons emitted when it falls from stationary state i to stationary state j.

The question immediately posed by this result is that if transitions between the same two stationary states involve photons of different energies, depending on whether the photon is emitted or absorbed, how can the observed resonance between identical atoms and molecules be explained?

As we have seen, the difference between the energy levels of a particle is neither an exact nor an absolute quantity; there is always an inherent uncertainty in its value.

Consequently, resonance between identical particles does not require the emission and absorption of identical photons.

So long as the difference between the energies of the emitted and absorbed photons, 2K, is less than the sum of the energy widths of the two states, $\Delta E = \Delta E_i + \Delta E_j$, resonance can occur, i.e., on condition that

$$\Delta E \geq 2K \tag{6.11}$$

The kinetic (recoil) energy, K, of the particle is related to its linear momentum, $p_{particle}$, by the equation

$$K = \frac{\left(p_{particle}\right)^2}{2m} \tag{6.12}$$

Substituting equation (6.7) in this expression gives:

$$K = \frac{E_{photon}^2}{2mc^2} \tag{6.13}$$

It follows, that there will be resonance between the particles so long as

$$\Delta E \geq \frac{E_{photon}^2}{mc^2} \tag{6.14}$$

6.2.3 Recoilless Emission and Absorption

The term mc^2 in equation (6.13) is equal to the rest-mass of the particle; for atoms and nuclei its value is of the order 10^9 to 10^{11}eV.

The energy of the photons emitted or absorbed by atoms in visible and ultra-violet transitions is between 1 and 10eV, and so the recoil energy,

$$K = \frac{E_{photon}^2}{2mc^2}$$

is of the order 10^{-11} to 10^{-7}eV.

The sum of the energy widths, ΔE, in such radiative transitions is usually substantially greater than this, and so the range of photon frequencies absorbed by an atom largely overlaps the range of frequencies emitted by identical atoms -see Figure 6.10a.

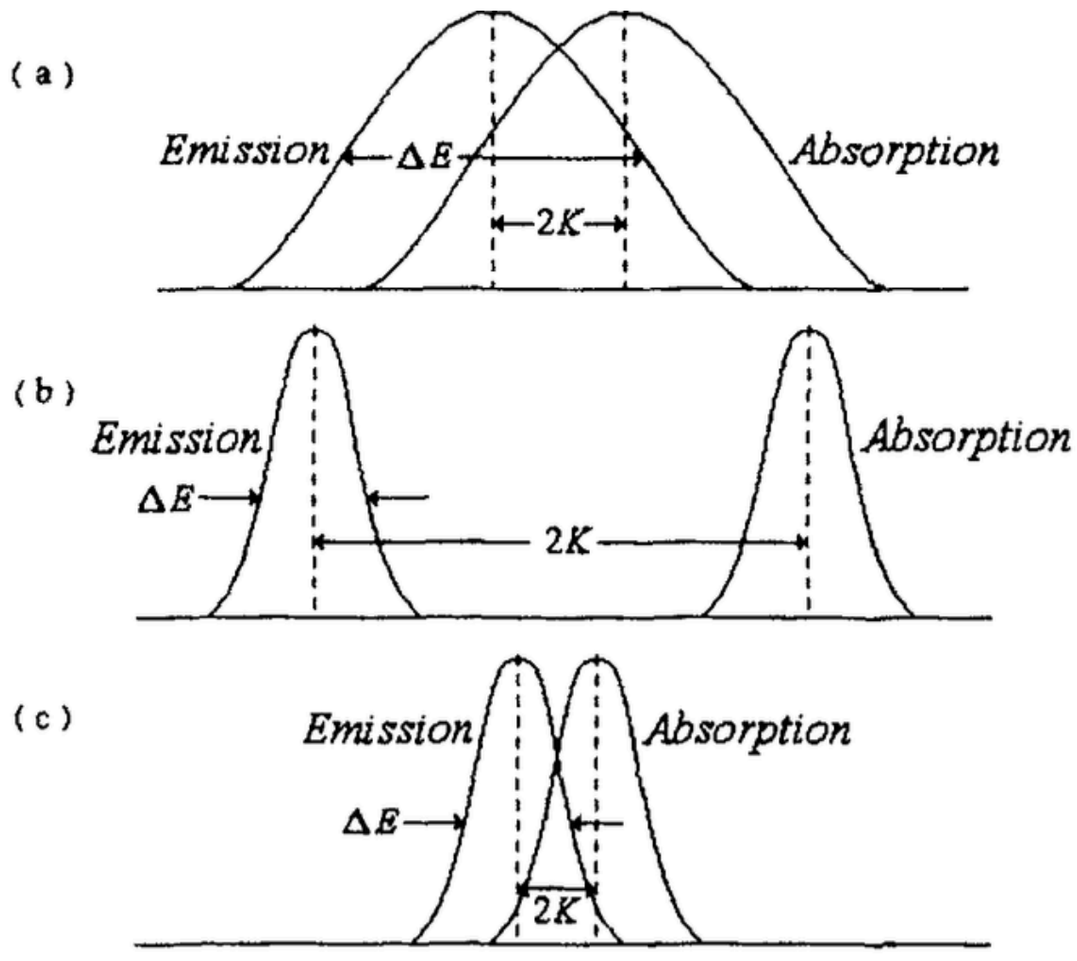


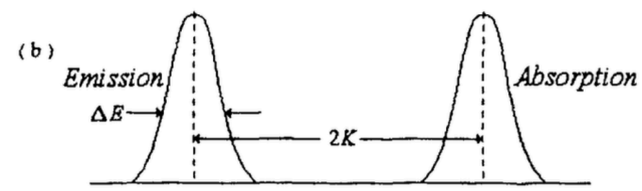
Fig.6.10 The inherent variance of the energy widths of two states in (a) a typical atomic or molecular system: the absorption and emission peaks overlap and there is resonance between the states; (b) a free nuclear system: the peaks do not overlap and there is no resonance; (c) a nuclear system in which the nuclei are embedded in a crystal: the peaks overlap and there is resonance.

In this case the recoil energy can be ignored, resonance conditions exist, and, within experimental limits, Bohr's equation is valid for transitions in both directions.

However, the energy, E, of the y ray photons emitted by radioactive nuclei is of the order 10^4 to 10^6eV: This gives a recoil energy, K, of the order 10^{-3} to 10^3eV: This amount of energy cannot be ignored; the sum of the energy widths, M, is much less than this.

Typical values for the lifetimes of excited nuclear states are 10^{-12} to 10^{-8}s, and so the natural width of the spectral lines is of the order 10^{-7} to 10^{-3}eV.

Consequently, the photons emitted by a radioactive nucleus are not absorbed by other identical nuclei; resonance conditions do not exist between the nuclei -see Figure 6.10b.



The conditions needed for γ ray resonance would be obtained, were it possible to reduce the recoil energy of the nucleus.

It is clear from equation (6.13) that the larger the particle mass m, the smaller the recoil energy K.

Radioactive nuclei that are embedded in a crystal structure do not behave, under certain conditions, as free particles.

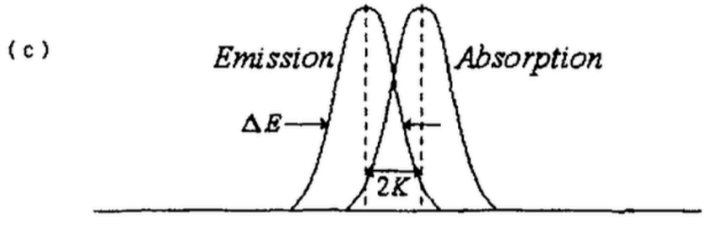
The particle mass, m, in this case is effectively equal to the mass of the crystal, which is about 10^{23} times that of a single nucleus.

Consequently, the sum of the energy widths, ΔE, will be much larger than the recoil energy.

The recoil energy of the crystal will be negligible, whatever the energy of the emitted or absorbed photon.

Thus, photons emitted by a radioactive nucleus embedded in a crystal will be absorbed by identical nuclei embedded in another crystal-see Figure 6.10c.

This is the phenomenon discovered by Mossbauer in 1958.



Because the nuclei are embedded in a crystal, their movement is restricted and there is no Doppler broadening of the spectral lines.

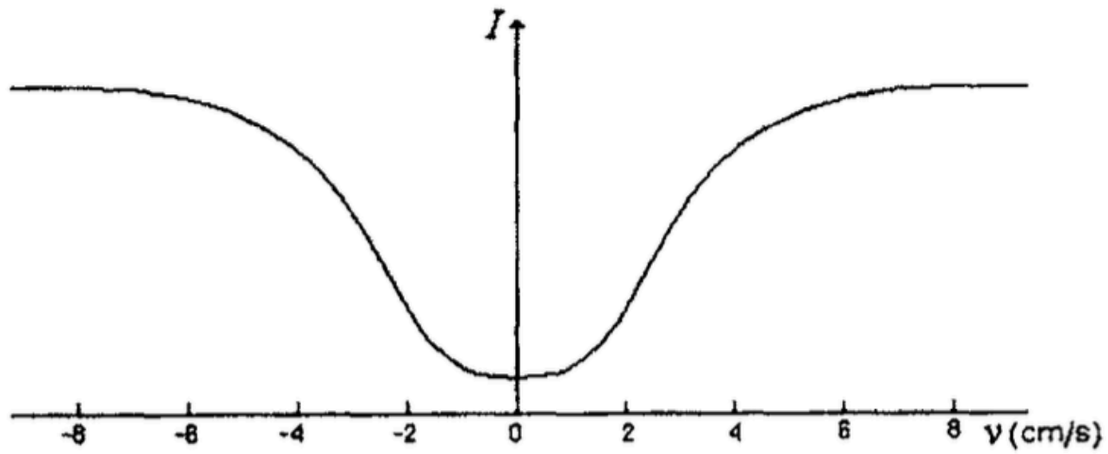
As a result, the width of the lines in the spectrum of Mossbauer radiation is just their natural width, that determined by the uncertainty principle.

The lines are very thin, and so resonance conditions are lost if there is even the slightest change in the radiation's frequency.

For example, should the emitting crystal be moving relative to the absorbing one, a Doppler change in the frequency of the radiation occurs.

Mossbauer found that a relative velocity of just 4 cm/s between the emitting and absorbing crystals was enough to cause the loss of resonance between them (Fig 6.11).

Fig.6.11 Typical results of an experiment using the Mossbauer effect. I is the intensity of the radiation transmitted by the absorber and v is the relative velocity between the emitter and the absorber. When the relative velocity is zero, almost all the radiation is absorbed. A relative velocity of just 4 cm/s is sufficient to destroy resonance conditions and allow transmission of the radiation through the absorber.



Both the energy width of the excited state in the radioactive nucleus and its average lifetime can be calculated from this result.

Nuclear lifetimes shorter than 10^{-11} seconds have been measured using the Mossbauer effect

Since any change in the energy levels of a nucleus affects the frequencies of the photons it emits and absorbs, the Mossbauer effect can be used to examine how changes in the immediate environment of a nucleus affect its energy levels.

For example, when a nuclear quantum state possesses a magnetic or electric moment that can interact with the internal electrical and magnetic fields of the crystal, the energy levels in the nucleus will be affected by its electrical and magnetic environment.

Consequently, identical nuclei, embedded in different crystals, will not necessarily be in resonance even when both crystals are at rest if there is a difference in their immediate environments.

The associated changes in the nuclear energy levels can be revealed and calibrated by determining the relative velocity between the crystals required to obtain resonance.

Alternatively, if the nuclear energy levels are known, the internal electrical and magnetic fields of the crystal can be studied, yielding information on its structure.

One of the principal nuclei employed for crystal research with the Mossbauer effect is the isotope of iron, Fe-57.

The energy of this isotope's first excited state is 14.4×10^3eV, and it has the relatively long average lifetime of 10^{-7}s. Its nuclear energy levels are affected by:

1. its valence state—whether the iron is bivalent or trivalent;
2. its neighbours in the crystal structure and by the crystal structure itself—whether, for instance, it is tetrahedral or octahedral;
3. whether the compound is ferromagnetic or paramagnetic.

Because the element iron is so widespread in nature and has an isotope that is particularly suited for Mossbauer effect experiments, it is widely used in the study of the solid state, and in chemistry, mineralogy and biology.

6.2.4 The Gravitational Shift - Black Holes

In 1960, Einstein's theory of general relativity was tested experimentally with the help of the Mossbauer effect.

We will describe the experiment's theoretical background and enumerate some of its repercussions.

Figure 6.12 illustrates a sequence of events which suggest, at first sight, that an infringement of mass-energy conservation might be a practical possibility.

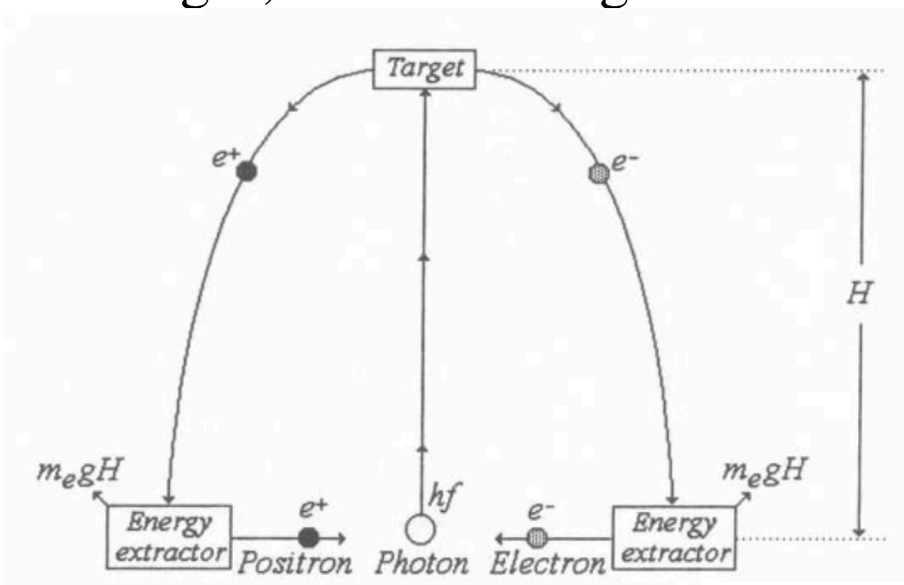


Fig.6.12 A photon of energy 1.02MeV strikes the target producing an electron-positron pair. These particles fall, meet and annihilate producing a new photon of energy 1.02MeV. In each cycle an amount $2m_egH$ of energy is usefully extracted.

A photon of energy 1.02MeV (twice the rest-mass energy equivalence, $2m_ec^2$, of an electron and a positron) is aimed upwards from the earth's surface.

At a height H, the photon strikes a target nucleus and an electron-positron pair is produced.

The electron and positron fall back to earth and in so doing acquire an amount of energy $2m_egH$ from the gravitational field.

This energy is extracted and used to perform useful work on the earth's surface.

After releasing this energy, the electron and positron collide, mutually annihilate and produce a new photon of energy 1.02MeV that is aimed upwards from the earth's surface restarting the cycle.

Thus, it would appear, that in each cycle an amount of energy, $\Delta E = 2megH$, can be usefully extracted from the process without any investment being made.

Clearly, if mass-energy is conserved, this is impossible.

The photon must lose an equal amount of energy during its rise from the earth's surface as the electron-positron pair gain during their fall if mass-energy is to be conserved.

This energy loss would be manifested as a fractional reduction, $\Delta f/f$, in the frequency of the radiation associated with the photon.

In terms of the energy, E, of the photon produced by their mutual annihilation, the mass, m_e, of the electron and of the positron is

$$m_e = E/2c^2 \tag{6.15}$$

and the amount of energy, ΔE, that must be added during each cycle for mass-energy conservation to be maintained is

$$\Delta E = 2m_egH = \frac{EgH}{c^2} \tag{6.16}$$

As a fraction of the photon's energy, E, this can be expressed as

$$\frac{\Delta E}{E} = \frac{gH}{c^2} \tag{6.17}$$

Substituting E = hf gives for the required fractional reduction in the photon's frequency

$$\frac{\Delta f}{f} = \frac{gH}{c^2} \tag{6.18}$$

The frequency of a photon that 'falls' a distance H should increase by the same extent.

The phenomenon is called the *gravitational shift*.

In an experiment that was carried out in 1960, the Mossbauer effect was employed to determine whether the frequency of a 'free-falling' photon does indeed change as predicted by equation (6.18).

In the experiment, γ-ray photons emitted from radioactive Fe-57 nuclei embedded in a stationary crystal 'fell' a distance of 22.5 meters.

The predicted fractional frequency change for this height was just 2.46 x10-^{15}.

However, the Mossbauer effect is so sensitive that even such small changes can be detected by it.

If indeed their frequency changed as predicted, the γ-ray photons would not be absorbed by an identical stationary crystal placed directly beneath the emitter; there would be no resonance between them.

However, if the absorbing crystal was itself receding from the emitter it would ascribe a lower frequency to the γ-ray photons due to a Doppler effect.

At the right speed, resonance would occur and the radiation would be absorbed.

In the experiment, the relative velocity between the emitting and absorbing crystals was adjusted until resonance was obtained, i.e., until the reduction in the photons' frequency due to the Doppler effect exactly cancelled out the increase in their frequency due to the gravitational field.

The Doppler change in the frequency of the photons was calculated from the velocity; it exactly matched that predicted by equation (6.18).

The shift observed in the frequency of the radiation can be explained in two different ways:

1. **In terms of gravitation: by ascribing a gravitational mass to the photons;**
2. **In terms of an accelerated motion: the consequence of a Doppler effect in an accelerated system.**

1. Assuming that the photons have a gravitational mass hf/c^2, the system's gravitational potential energy, U, will change when a photon rises or falls a short distance H by an amount:

$$\Delta U = m_{photon} gH = \left(\frac{hf}{c^2}\right) gH \tag{6.19}$$

This change in gravitational potential energy is matched by a corresponding change in the photon's energy, $\Delta E_{photon} = h(f - f')$, when its frequency changes from f to f'. Equating the energy changes gives:

$$h(f - f') = \left(\frac{hf}{c^2}\right) gH \tag{6.20}$$

from which we obtain the expression

$$\frac{f - f'}{f} = \frac{\Delta f}{f} = \frac{gH}{c^2} \tag{6.21}$$

which is identical to equation (6.17).

2. **Alternatively, suppose that the experiment illustrated in Figure 6.12 is conducted in a free-falling laboratory close to the earth's surface; a gravityless laboratory. Moving with a velocity c, the photon strikes the target at a time $\Delta t \approx H/c$ after its upward release. During this time the velocity, v, of the laboratory increases by an amount:**

$$\Delta v = g\Delta t = gH/c \qquad (6.22)$$

As a result of this change in velocity, there will be a Doppler change in the frequency of the photon:

$$\frac{\Delta f}{f} = \frac{\Delta v}{c} = \frac{gH}{c^2} \qquad (6.23)$$

This relationship is also identical to equation (6.17)

Hence, the frequency shift can be explained as being the result either of a gravitational force or of an acceleration, as required by the principle of equivalence.

Accordingly, this experiment is regarded as giving confirmation to the theory of general relativity.

The frequency of a photon that 'rises' in a gravitational field decreases; this effect is called a *red-shift*.

The *gravitational red-shift* of the radiation emitted by galaxies and stars is usually masked by the greater Doppler shift in its frequency attributed to their recessional motion.

Nevertheless, we can imagine a body whose gravitational field is so strong that a photon has insufficient energy to escape from it; such a body is called a *black hole*.

It has been conjectured that black holes could be formed from the complete gravitational collapse of the supernova core that remains after the explosive demise of a massive star.

The boundary of a black hole, known as its *event horizon*, is the envelope formed by the space-time paths of the light rays that just fail to escape from it.

Hawkings has suggested that despite the tremendous gravitational force exerted by a black-hole, mass-energy in the form of radiation and particles should nevertheless be able to leak from it.

His hypothesis, which is based on general relativity and quantum theory, can be understood as follows:

One of the corollaries of the uncertainty principle is that empty space cannot be a complete void.

If it were, the value of the electromagnetic and gravitational fields in the empty space would be precisely zero, in contravention of the uncertainty principle.

Empty space must be regarded as a fluctuating field of particle pairs, that are continuously appearing and disappearing; particles being constantly created and annihilated.

A particle and its anti-particle form, move apart, come together again a short time later and finally mutually annihilate, the process being repeated unceasingly throughout the expanse of 'empty' space-time.

Suppose, that as a result of these *vacuum fluctuations*, a particle pair forms just beyond the event horizon of a black hole and that, before they manage to converge and mutually annihilate, one of them 'falls' into the black hole.

Bereft of its partner, the particle remaining outside the event horizon has little prospect of annihilation; it becomes destined to exist.

Alone, it moves away from the event horizon and to an observer at some distance from the black hole it appears as though the particle was emitted from it.

The particle possesses mass-energy that can only be accounted for as having been acquired at the expense of the black hole.

Thus, black holes appear gradually to lose mass-energy, ever shrinking as they radiate more and more particles until they finally disappear in one last cataclysmic explosion.

Part 6.3 Nuclear Magnetic Resonance

The concept of nuclear magnetism dates back almost to the time of Rutherford's original discovery of the atomic nucleus.

The components of the nucleus were initially thought to be protons, electrons and perhaps ex. particles, and it was presumed that these charged particles would be involved in some sort of rotational motion and would therefore give rise to a magnetic moment.

In 1924, Pauli suggested that the hyperfine structure of atomic spectra resulted from an interaction between the orbital motion of the electrons and the angular momentum of the nucleus.

Later researchers suggested that the interaction might be magnetic and the first values of nuclear magnetic moments were inferred from spectral measurements.

Initially, precise measurements of nuclear magnetic moments proved hard to achieve; they are some three orders of magnitude smaller that that of the electron.

The breakthrough in accurate measurements came when techniques were discovered for inducing the transition of the magnets associated with nuclear spin from one energy level to another.

Under the influence of an oscillating magnetic field, the nuclear spins resonantly absorb energy from the magnetic field or release energy to the field by stimulated emission.

This can be detected by suitable instrumentation.

At first, these nuclear magnetic phenomena, termed *nuclear magnetic tesonance - NMR,* were of interest mainly to physicists.

But when, in 1951, it was reported that the magnetic moments associated with the protons in ethanol (CH3CH20H) absorb/emit energy at three different frequencies and at intensities in the ratio of 3:2:1, nuclear magnetic resonance had arrived for the organic chemist.

Since then, it has established itself as one of the most valuable tools for studying chemical compounds. Few analytic or diagnostic techniques provide such an abundance of accurate data.

In many cases the nucleus under investigation is that of a hydrogen atom, i.e., a proton; the effect is then often referred to as *proton magnetic resonance* - PMR.

A similar mechanism is involved in *electron spin resonance* - ESR where the magnetic dipole concerned is that associated with an electron's angular motion.

For an understanding of how nuclear magnetic resonance works, we must start by examining the sources of the magnetic properties of nuclei.

6.3.1 Magnetism and Angular Momentum

Despite the repeated attempts that have been made to isolate a natural unit of magnetism similar to the natural unit of electricity-a magnetic monopole or charge one has never been discovered or identified.

Magnetic poles always seem to occur in pairs; a north pole facing one way and a south pole facing the other.

Thus. the fundamental ‘particle’ of magnetism appears to be the *magnetic moment or magnetic dipole* whose existence is attributed to the angular motion of charged elementary particles.

Examples of this angular motion are:

- **the motion of an electron around a conducting loop;**
- **the orbital motion of an atomic electron relative to the nucleus;**
- **the intrinsic spin of an electron or of any other charge carrying elementary particle.**

In general, the magnetic dipole moment, $\boldsymbol{\mu}$, associated with a particle's angular momentum, **J**, is given by:

$$\boldsymbol{\mu} = \gamma \mathbf{J}\left(\frac{e}{2m}\right) \tag{6.24}$$

where m is the particle's mass, e is the natural unit of charge and γ a parameter of the system called the *gyromagnetic ratio*.

For example, the magnetic moment associated with the orbital motion of a particle is given by:

$$\boldsymbol{\mu} = \gamma_L \mathbf{L}\left(\frac{e}{2m}\right) \tag{6.25}$$

where **L** is the particle's orbital angular momentum.

In the case of a single electron, the orbital gyromagnetic ratio $(\gamma_L)_{electron} = -1$.

For an orbiting proton, the gyromagnetic ratio $(\gamma_L)_{proton} = +1$.

The magnetic moment associated with a particle's spin, **S**, is given by:

$$\boldsymbol{\mu}_S = \gamma_S \mathbf{S}\left(\frac{e}{2m}\right) \tag{6.26}$$

where γ_S is the particle's spin gyromagnetic ratio. The spin gyromagnetic ratios of the electron, proton and neutron are respectively:

$$(\gamma_S)_{electron} = -2.0024; \quad (\gamma_S)_{proton} = +5.5851; \quad (\gamma_S)_{neutron} = -3.8256.$$

In general, the energy possessed by a magnetic dipole situated in a magnetic field is given by the scalar multiplication:

$$E = -\boldsymbol{\mu} \cdot \mathbf{B} = -\mu B \cos\theta \tag{6.27}$$

where θ is the angle between the direction of the magnetic dipole vector, μ, and that of the magnetic field vector, **B**.

The dipole has its minimum energy, $E_{min} = -\mu B$, when it is aligned with the direction of the field, i.e., when $\theta = 0°$, and its maximum energy, $E_{max} = +\mu B$, when its direction is opposite to that of the field, i.e., when $\theta = 180°$.

6.3.2 Nuclear Magnetic Moments

Atomic nuclei possess an intrinsic angular momentum, **I**, called *nuclear spin*, that arises from the combination of the intrinsic angular momentum (spin) of their constituent nucleons with their motion inside the nucleus.

Nuclear spin is quantized. The eigenvalues of the square, I^2, of the magnitude of the nuclear spin are given by:

$$I^2 = I(I+1)\hbar^2 \tag{6.28}$$

where I is the *nuclear spin quantum number*.

Relative to any arbitrarily chosen axis, the nuclear spin can take up $(2I + 1)$ orientations. The permitted values of the component, I_z, of nuclear spin parallel to an arbitrarily chosen Z-axis are:

$$I_z = m_I \hbar$$
$$\text{where } m_I = -I, (-I+1), (-I+2) \ldots, +I \tag{6.29}$$

It has been found, empirically, that the values of the nuclear spin quantum number are subject to the following rules:

- **In nuclei containing an even number of nucleons, the nuclear spin quantum number, *I*, is an integer.**
- **In nuclei containing an odd number of nucleons the nuclear spin quantum number, *I*, is an integer multiple of a half.**
- **In almost all nuclei containing an even number of both protons and neutrons, the nuclear spin quantum number $I = 0$.**

These rules suggest that the protons and neutrons in a nucleus each tend to pair their angular momenta in opposite directions.

By convention, the nuclear magnetic moments associated with nuclear spin are expressed in terms of the mass of the proton, m_p,

Accordingly, the magnetic moment of a nucleus is given by

$$\boldsymbol{\mu}_{Nucleus} = \gamma_N \mathbf{I}\left(\frac{e}{2m_p}\right) \tag{6.30}$$

where γ_N is the *nuclear gyromagnetic ratio*.

The allowed components, $(\mu_{Nucleus})_z$ of the nuclear magnetic moment parallel to an arbitrarily chosen Z-axis are given by:

$$\left(\mu_{Nucleus}\right)_z = \gamma_N I_z\left(\frac{e}{2m_p}\right) = \gamma_N m_I \hbar\left(\frac{e}{2m_p}\right) \tag{6.31}$$

This equation can be written in the concise form

$$\left(\mu_{Nucleus}\right)_z = \gamma_N \mu_N m_I \qquad (6.32)$$

$$\text{where } m_I = -I, \left(-I+1\right), \left(-I+2\right) \ldots, +I$$

The quantity μ_N is called the *nuclear magneton*:

$$\mu_N = \frac{e\hbar}{2m_p} = 5.051 \cdot 10^{-27}\,\mathrm{J \cdot T^{-1}} \qquad (6.33)$$

Its value is 1836 times smaller than the Bohr magneton.

In a magnetic field, **B**, the energy levels, E_{m_I} associated with the z-component of the nuclear magnetic moment are given by:

$$\begin{aligned} E_{m_I} &= -\left(\mu_{Nucleus}\right)_z B_z \\ &= -\gamma_N \mu_N m_I B_z \qquad (6.34) \\ \text{where } m_I &= -I, \left(-I+1\right), \left(-I+2\right) \ldots, +I \end{aligned}$$

In the case of a hydrogen nucleus (a proton), whose nuclear spin quantum number is $I = 1/2$, just two energy levels are possible.

The lower level, $E_{+1/2}$, corresponds to the $m_I = +1/2$ state and the higher, $E_{-1/2}$, to the $m_I = -1/2$ state.

The difference between the two energy levels is

$$\begin{aligned}\Delta E &= -\gamma_N \mu_N B_z\left[-\tfrac{1}{2} - \left(+\tfrac{1}{2}\right)\right] \\ &= 5.5851 \cdot 5.051 \cdot 10^{-27} B_z \\ &= 28.2 \cdot 10^{-27} B_z\end{aligned} \tag{6.35}$$

Note that the difference between the energy levels is a linear function of B_z, the magnitude of the z-component of the magnetic field (Fig 6.13).

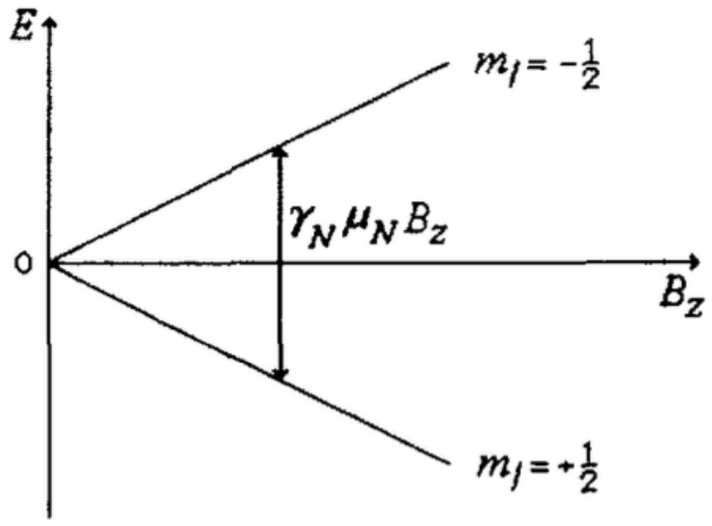


Fig 6.13 Diagram of the energy levels of the z-component of the magnetic moment of a proton in an external magnetic field. The stronger the magnetic field, the greater the difference between the levels.

Because the nuclear magneton, μ_N, is some three orders of magnitude smaller than the Bohr magneton, μ_B, the differences between the energy levels associated with nuclear magnetic moments are substantially lower than those associated with electronic magnetic moments.

For a proton immersed in a magnetic field of magnitude $B_z = 1T$, the difference, ΔE, between its two spin energy levels is

$$\Delta E = \frac{28.2 \cdot 10^{-27}}{1.6 \cdot 10^{-19}} \cdot 1 = 17.63 \cdot 10^{-8}\,\text{eV} \tag{6.36}$$

This is a small amount compared to the thermal energy of the atoms.

Hence, given the Maxwell-Boltzmann distribution, the ratio between the populations of the two spin states at room temperature (~300K) will be:

$$\begin{aligned} N_{Lower}/N_{Higher} &= e^{\Delta E/kT} \\ &= e^{\left(17.63\cdot10^{-8}/8.62\cdot10^{-5}\cdot300\right)} \\ &= 1.000007 \end{aligned} \tag{6.37}$$

i.e., the population of the lower state is only very slightly greater than that of the higher state.

The ratio can be increased by cooling the system or by increasing the magnitude of the magnetic field.

The smallness of the nuclear magneton also affects the transition frequencies between the nuclear spin energy levels.

In radiative transitions between the proton spin energy levels in a magnetic field of magnitude $B_z = 1$ T, a photon of frequency

$$f = \frac{\Delta E}{h} = \frac{17.63 \cdot 10^{-8}}{4.14 \cdot 10^{-15}} = 4.26 \cdot 10^{7}\,\text{Hz} \tag{6.38}$$

will be emitted or absorbed. This frequency is in the radio region of the electromagnetic spectrum.

By contrast, radiative transitions associated with electronic spin energy levels are in the microwave region.

Typical values for nuclear magnetic data are given in Table 6.1.

Table 6.1: Magnetic Data of Selected Nuclides

Nuclide	Nuclear Spin Quantum Number, I	Transition Frequency in a Magnetic Field of Magnitude 1.4094T[18]	Nuclear Gyromagnetic Ratio - γ_N
^{1}H	½	60.000 MHz	5.585
^{2}D	1	9.210	0.857
^{7}Li	1½	23.317	2.171
^{14}N	1	4.334	0.403
^{15}N	½	6.080	–0.567
^{17}O	2½	8.134	–0.757
^{19}F	½	56.446	5.257
^{23}Na	1½	15.871	1.478

6.3.3 Nuclear Magnetic Resonance

In semi-classical terms, the behavior of a nuclear magnetic moment in a magnetic field is analogous to that of the orbiting electron in the Bohr-Sommerfeld model of the hydrogen atom.

It can, likewise, be visualized as precessing around the direction of the field (Fig 6.14).

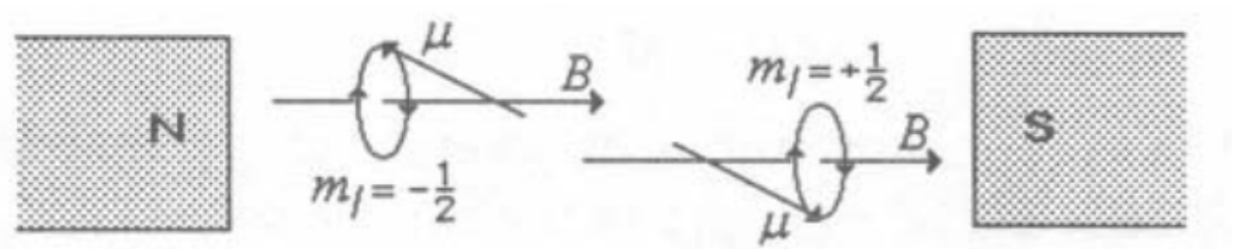


Fig 6.14 Semi-classical visualization of the precession of a proton's magnetic dipole moment, μ, around the direction of a uniform magnetic field, B. The diagram shows the two possible orientations of the dipole, one corresponding to $m_I = +1/2$, the other to $m_I = -1/2$.

On this view, the transition frequency of the nuclear magnetic dipole is the Larmor frequency of its precessional motion, f_L, and the photons emitted or absorbed in these transitions will be of energy hf_L.

Since the Larmor frequency is itself a function of the applied magnetic field, B, the frequency of the photons emitted or absorbed will also depend on the field strength.

At the radio frequencies involved in magnetic dipole transitions, the probability of the spontaneous emission of a photon is negligible.

Thus, both the absorption and emission of photons from nuclear magnetic dipoles requires the stimulus of an external source.

On the semi-classical model, the absorption or emission of a photon by a nuclear magnetic dipole involves a change in its orientation relative to the applied field, B.

In the case of a proton, where $I = 1/2$, this requires flipping the magnetic dipole through 180°, i.e., from $m_I = +1/2$ to $m_I = -1/2$, or vice versa.

In order to flip the precessing magnetic dipoles through 180°, a force must be applied to them at right angles to the axis around which they are precessing.

This can be achieved by the action of a secondary oscillating magnetic field, B', whose frequency equals the Larmor frequency of the dipoles' precessional motion.

Only when the oscillations of the secondary 'tipping' field are exactly synchronized with the precessional motion of the dipole is the transition induced and is energy exchanged between the field and the dipole (Fig 6.15).

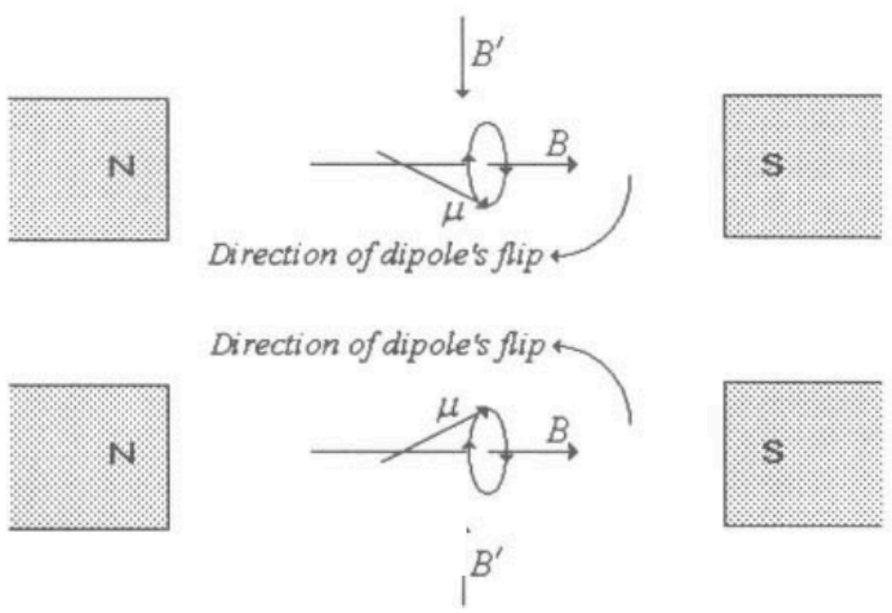


Fig 6.15 Nuclear Magnetic Resonance. Only when the frequency of the oscillating secondary magnetic field, B', equals that of the dipole's precessional motion, does a turning moment that tends to flip the axis of its precession through 180° act continuously on the dipole. Under these conditions, the dipole can flip from the $m_I = +1/2$ state to the $m_I = -1/2$ state.

6.3.4 Observing Nuclear Magnetic Resonance

Nuclear magnetic resonance can be detected by several techniques.

We will describe three of them: Molecular-beam Resonance, Absorption Resonance and Induction Resonance.

In all of these, the material whose nuclear magnetic dipole moments are under examination, is immersed in a uniform magnetic field, B, that is crossed by an oscillating magnetic field, B' (Fig 6.16).

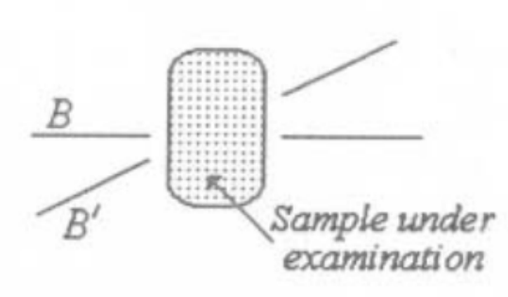


Fig 6.16 The crossed uniform magnetic field, B, and oscillating magnetic field, B'.

Although, in principle, resonance can be located either by adjusting the frequency of the oscillating field, B', to that of the dipoles' precessional (Larmor) frequency or by varying the magnitude of the uniform field, B, until the frequency of the dipoles' precession equals that of the oscillating magnetic field, in practice, the latter option is usually technically easier.

The three methods we describe differ in the way they detect that resonance has been achieved between the precessing nuclear dipole and the oscillating magnetic field.

1. Molecular-beam Resonance

This technique, which is closely related to the design of the Stem-Gerlach experiment, was first developed by Rabi.

It likewise exploits the effect of the non-uniform magnetic field produced in the space between the differently shaped poles of an electromagnet on spin magnetic moments.

Gaseous molecules emitted by a heated source are introduced into one end of the space between the differently shaped poles of an electromagnet (Fig 6.17).

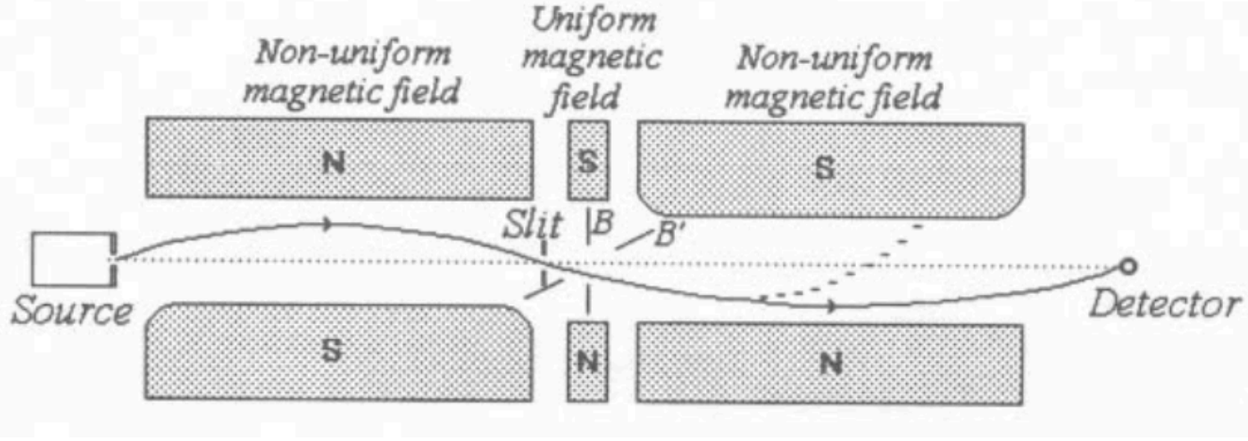


Fig 6.17 Molecular-beam Resonance. The oscillating magnetic field B' is generated by passing a radio frequency alternating current through a coil. If the orientation of the nuclear magneti dipole moments is unaffected by the crossed magnetic fields, B and B', the molecular beam reaches the detector. However, if a change occurs in their orientation the path taken by the molecules through the second non-uniform magnetic field (the broken line) will not match the one they took through the first field and so they will not reach the detector.

The molecules emerging from the slit at the far end of this space enter a narrow region in which a uniform magnetic field, B, is crossed by an oscillating magnetic field, B'.

The molecules enter the region at right angles to both fields.

If the orientation of their nuclear magnetic dipole moments is unchanged during the passage of the molecules through the crossed fields, any deflections they had sustained in their passage through the non-uniform magnetic field can be annulled by passing them through a second identical but reversed non-uniform field.

However, should there be resonance between the nuclear magnetic dipoles and the oscillating magnetic field, they will flip into new orientations.

In these new orientations, the deflections they had sustained in the first non-uniform magnetic field will not be annulled by passing them through an identical but reversed second non-uniform magnetic field.

Resonance between the dipoles and the oscillating field, B', will be indicated by a reduction in the beam intensity registered by a detector positioned at the exit from the second non-uniform field.

In the experiments carried out by Rabi and his co-workers, the frequency of the oscillating magnetic field, B', was held constant as the magnitude of the uniform field B was varied.

Typical results they obtained are shown in Fig 6.18.

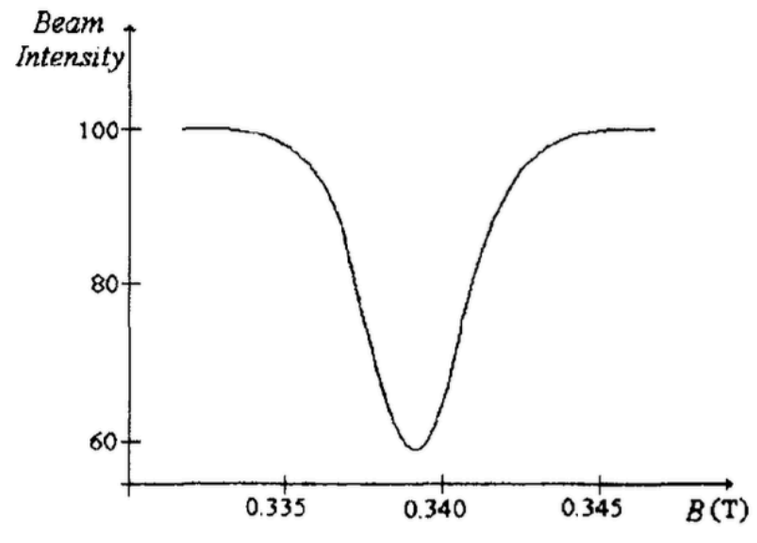


Fig 6.18 The beam intensity as a function of the magnitude of the uniform field, B, in an experiment with LiCl molecules. The frequency of the oscillating field was 5.585MHz. Resonance due to the ^{7}Li isotope occurred at B = 0.3384T.

The molecular-beam technique can only be employed with gases.

Nuclear magnetic resonance in liquids and solids cannot be investigated by this method.

The following two methods were subsequently developed to answer this need.

2. Nuclear Absorption Resonance

This technique for observing nuclear magnetic resonance in liquids and solids was developed by Purcell.

It involves measuring the nuclear absorption of the energy carried by radio waves whose oscillating magnetic field supplies the 'tipping' B' field (Fig6.19).

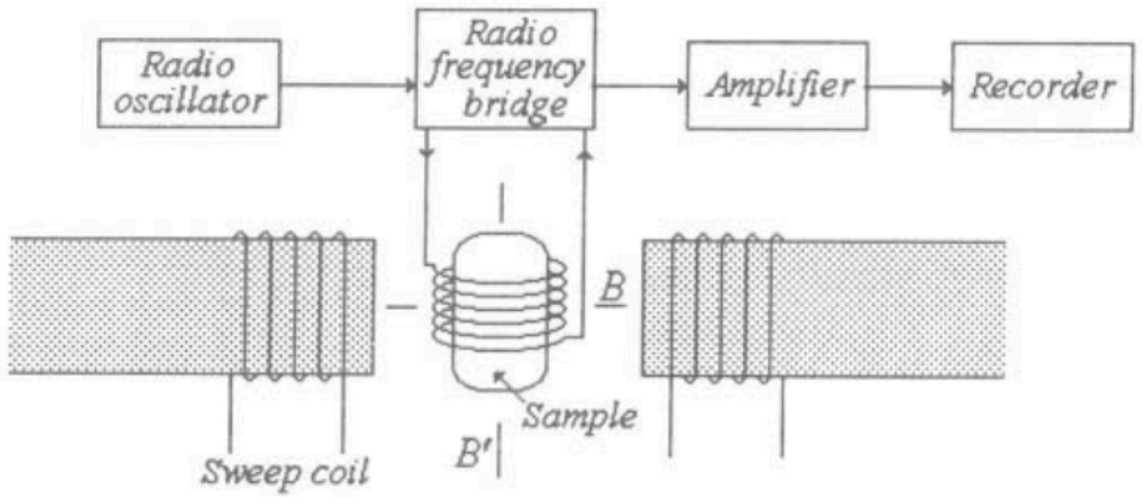


Fig 6.19 Schematic diagram of nuclear absorption NMR spectrometer. The frequency of the secondary field, B', is held constant. The primary field, B, is varied by means of the sweep coil. Changes in the intensity of the radio waves are recorded as a function of the primary field B.

The material under examination is irradiated with radio waves of a fixed frequency.

At resonance, nuclei in the sample absorb energy from the oscillating field.

This is indicated by a sharp attenuation of the intensity of the radio waves.

The recorded spectrum shows the intensity of the radio waves as a function of the magnitude of the primary field (Fig 6.19).

Absorption peaks correspond to resonance between the field and the nuclei.

Only those nuclei that are in the lower spin state can absorb energy from the field.

Thus, the strength of the radio frequency signal picked up by the spectrometer depends on the number of nuclei rising from the lower to the higher state.

However, as equation (6.37) shows, the population of the ground state is only marginally greater than that of the excited state.

The situation can be improved by cooling or by using very strong magnets to produce the B-field-up to l0 T.

However, maintaining the uniformity of such fields is a great technical challenge.

A further problem is that if the sample is irradiated by a continuous strong oscillating field, the populations of the two states rapidly equalize.

With equal populations in each state, the number of photons absorbed by the sample will equal the number emitted and no signal will be received.

This can be avoided either by employing weak oscillating fields or, alternatively, by using strong fields in short pulses, between which the populations of the two states can return to their equilibrium numbers.

3. Nuclear Induction Resonance

This technique for observing nuclear magnetic resonance in liquids and solids was developed by Bloch.

It involves detecting the radiation emitted by the excited nuclei as they fall from a higher spin state to a lower one (Fig 6.20).

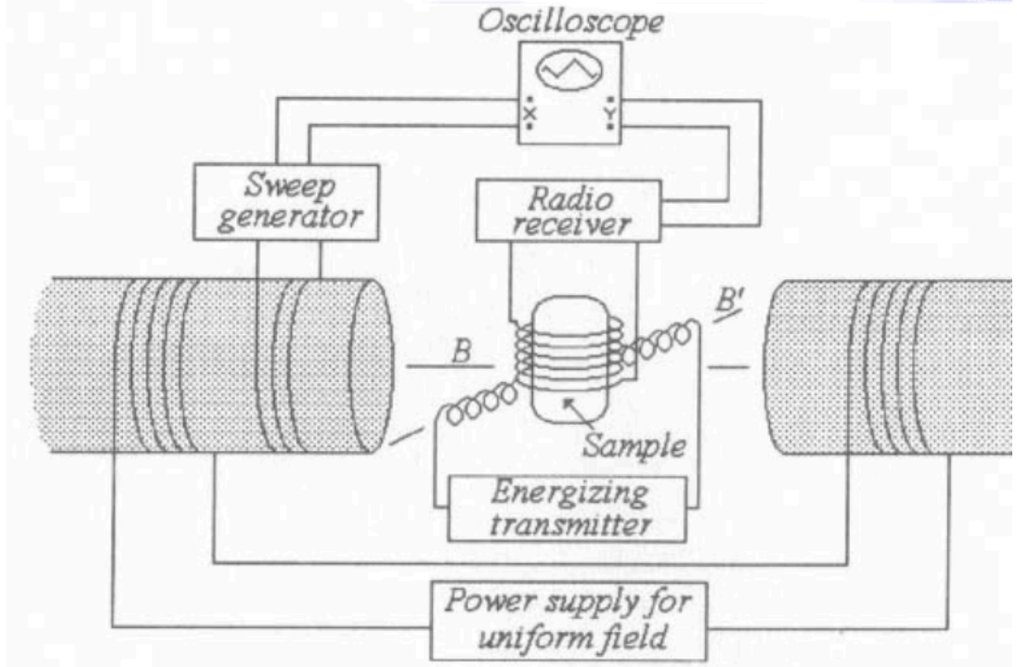


Fig 6.20 Schematic diagram of nuclear induction NMR spectrometer. The frequency of the secondary field, B', induced by the energizing transmitter is held constant. The magnitude of the primary field, B, is varied by means of the sweep coil. Changes in the intensity of the radio waves emitted are recorded as a function of the primary field B. A computer can replace the oscilloscope.

The radio frequency radiation is picked up by a coil (antenna) whose axis is at right angles to both the uniform filed, B, and the oscillating field, B'.

After amplification, the signal from the coil is typically fed into the Y-plates of an oscilloscope whose X-plates are connected to the sweep coil that varies the magnitude of the uniform field, B.

At resonance, nuclei are excited into a higher spin state by the oscillating field, B', generated by the energizing transmitter and the intensity of the radiation they subsequently emit increases correspondingly.

The spectrum appears on the oscilloscope screen as a trace showing the intensity of the emitted radiation as a function of the magnitude of the uniform field, B (Fig 6.21).

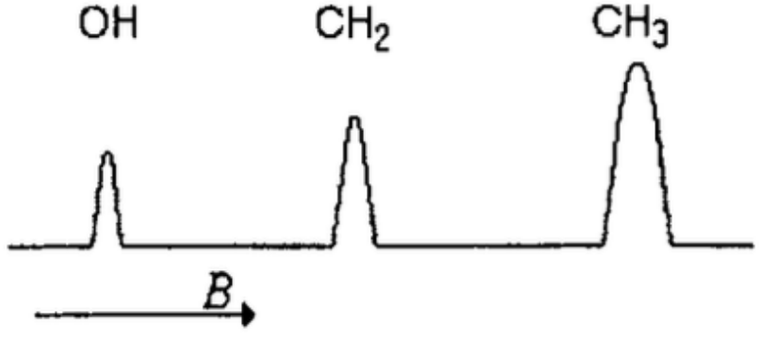


Fig 6.21 The low resolution NMR spectrum of ethanol.

In an adaptation of the induction spectrometer, the sample is irradiated with a range of radio frequencies -white radiation.

In order to detect the dominant frequencies in the ensuing radiation emitted by the sample, it is subjected to a Fourier transform.

This technique is extremely sensitive and can be used to study the NMR spectrum of nuclei present in low concentrations, such as ^{13}C.

6.3.5 Chemical Shift

The effective magnetic field, **B**., that actually determines the energy levels of the nuclear magnetic moments in the sample under examination in NMR absorption or induction spectroscopy differs from the applied B-field.

Under the influence of the B-field, the atomic electrons in the material are set into motion.

This motion in tum induces a secondary magnetic field, $\mathbf{B}_{ind}$.

The effective field experienced by the nuclear magnetic moments is the resultant of the applied field, **B**, and the induced field, $\mathbf{B}_{ind}$:

$$\mathbf{B}_{Eff} = \mathbf{B} + \mathbf{B}_{Ind} \tag{6.39}$$

The induced field generally opposes the applied B-field.

This is an instance of the familiar phenomenon of diamagnetism discovered by Faraday.

The nature of the induced secondary field, $\mathbf{B}_{ind}$, depends on the electronic configuration of the system. In general,

$$\mathbf{B}_{Ind} = \sigma\mathbf{B} \tag{6.40a}$$

where σ is a constant characteristic of the nuclear environment called the shielding constant.

It follows, that the effective field experienced by the nuclear magnetic moment is

$$\mathbf{B}_{Eff} = \mathbf{B}(1 - \sigma) \tag{6.40b}$$

The frequency of the photons absorbed and emitted by a nuclear magnetic moment depends on the actual magnetic field it experiences, which in tum depends on its electronic environment.

Thus, any change in the immediate electronic environment of the nuclear magnetic moments will result in a shift in the photon frequency.

This is the basis for the widespread use of NMR spectroscopy in the study of molecular structure.

For a given nuclear magnetic dipole moment, say that of a proton, the greater the diamagnetic shielding it experiences, the stronger the external B-field required for resonance to be observed.

The low resolution NMR spectrum of ethanol, CH_3CH_2OH shown in Fig 6.21 is an example of this effect.

Three distinct bands appear in the spectrum corresponding to the CH_3, CH_2 and OH groups.

The area of each band is proportional to the number of hydrogen atoms each comprises.

It is apparent that a stronger field is required for the resonance of the CH_3 protons than for the CH_2 protons, indicating the former's greater diamagnetic shielding; the proton in the OH group is the least shielded.

Because the B-field magnitude at which resonance occurs depends on the frequency of the oscillating B'-field, it is meaningless to express NMR data in absolute units; a value of the B-field magnitude that gives rise to resonance in one spectrometer will not necessarily produce resonance in another spectrometer that operates at a different frequency.

For this reason, NMR data are always expressed with reference to a standard substance.

A standard commonly employed in proton magnetic resonance is Tetramethylsilane or TMS, $Si(CH_3)_4$, which contains 12 magnetically identical protons.

The shielding constant of TMS is relatively large and its NMR spectrum is a single sharp line.

Moreover, TMS is chemically inert and miscible with many solvents and so can be added as an internal standard to the samples under investigation without affecting them chemically or magnetically.

Suppose, that in a particular NMR spectrometer, the protons in TMS resonate at a field magnitude of B_{TMS} and those in the CH_3 group in ethanol at a field magnitude of B_{CH_3}.

This will be reported in the form

$$\delta_{CH_3} = \frac{B_{TMS} - B_{CH_3}}{B_{TMS}} \cdot 10^6 \text{ parts per million} \qquad (6.41)$$

where δ_{CH_3} is called the *chemical shift* of the protons comprising the CH_3 group.

The NMR spectrum of ethanol exhibits three chemical shifts: that corresponding to the three protons in the CH_3 group, that of the two protons in the CH_2 group and that of the single proton in the OH group.

All the protons in each group share the same chemical environment.

The range of chemical shifts for most hydrogen nuclei (protons) is about ten parts per million.

Chemical shifts can also be expressed in terms of frequencies:

$$\delta_{Nucleus} = \frac{f_{Nucleus} - f_{\mathrm{TMS}}}{f_{Spectrometer}} \cdot 10^6 \text{ parts per million} \tag{6.42}$$

where $\delta_{\mathrm{Nucleus}}$ is the resonance frequency of the magnetic moment of the nucleus under investigation, f_{TMS} is the resonance frequency of the protons in the TMS molecule and $f_{\mathrm{spectrometer}}$ is the operating frequency of the NMR spectrometer.

The secondary magnetic fields responsible for the chemical shifts are very weak, typically less than 0.001% of the applied B-field.

In a B-field of magnitude 1.4094T, the maximum displacement in the resonance frequencies of hydrogen nuclei (protons) due to their chemical environment is only 600Hz.

This is a very small shift compared to the resonance frequency of free protons in such a field, viz. 60MHz.

Thus, the detection of chemical shifts requires great experimental precision. In order to achieve a resolution of ±lHz, the uniformity of the B-field and the frequency of the B'-field must vary by no more than one part in sixty million.

Higher resolutions require still greater precision.

The ultimate limit on the resolution that can be obtained by any spectroscopic technique is set by the inherent natural width of the spectral lines.

From Heisenberg's uncertainty principle, $\Delta E\, \Delta t \approx \hbar$, and the relationship $\Delta E = h\, \Delta f$, the natural width, Δf, of a spectral line is given by

$$\Delta f \approx 1/(2\pi \Delta t) \tag{6.43}$$

where Δt is the lifetime of the excited state. The lifetimes of nuclear spin states are often called their *relaxation times*.

Nuclear relaxation times are affected by the coupling that occurs between the nuclear magnetic moments.

Relaxation times in solids are short, $\sim 1\mu s$, because of the direct spin-spin dipole coupling that occurs between their magnetic moments.

As a result, the NMR. spectral lines of solids are broad, $\Delta f \approx 100{,}000\text{Hz}$, and can provide no information about chemical shifts.

In certain cases, the lines can be narrowed significantly by spinning the solid sample at high speed at so-called magic angles relative to the direction of the B-field.

On the other hand, because of the rapid random motion of their molecules, the coupling of the nuclear magnetic moments in liquids is negligible.

Thus, relaxation times in liquids are relatively long, ~ ls, and the NMR spectral lines are correspondingly narrow, ~ 0.l5Hz.

For example, the inherent uncertainty of the standard 60MHz/1.4904T proton line is less than one part in a hundred million.

To take advantage of this inherent accuracy and the opportunity it presents for gaining information about the proton and its close magnetic environment, the magnetic field, B, in both the absorption and induction NMR spectrometers must be uniform to a degree of one part in a hundred million.

To average out any remaining field inhomogeneities, the sample tube is usually spun at about 60 revolutions per minute about an axis perpendicular to the B-field.

The variation in the frequency of the B'-field must also be less than one part per hundred million.

The high resolution NMR. spectrum of ethanol shows that the CH_2 and CH_3 bands are split into a number of narrower bands (Fig 6.22).

The phenomenon, called spin-spin coupling, results from the electron mediated interactions of the various nuclear spins in the molecule.

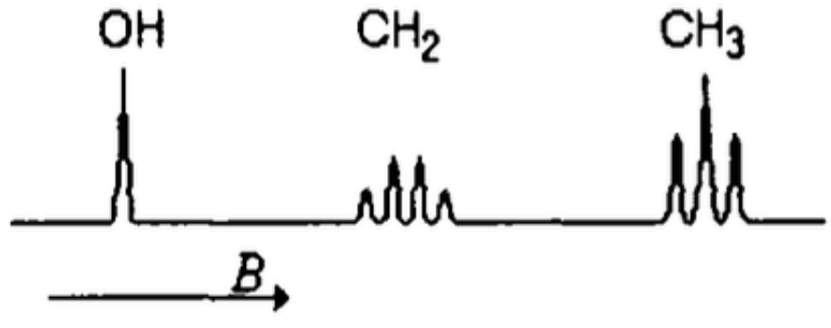
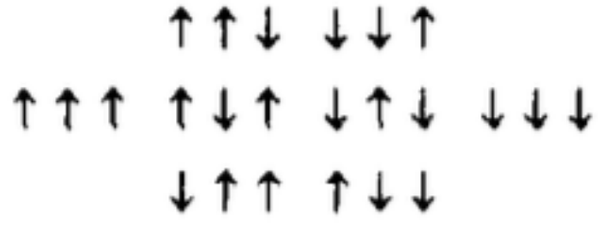


Fig 6.22 The high resolution NMR spectrum of ethanol.

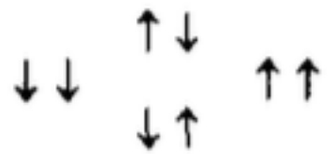
Because of its mobility, the proton in the hydroxyl group does not couple with the other protons in the molecule.

However, because the nuclear spins of the three protons in the CH_3 group can be arranged in four distinct ways,



each of the two protons in the CH_2 group can couple with this neighbor in four ways; the result is the splitting of its spectral line into a quartet.

On the other hand, because the nuclear spins of the two protons in the CH_2 group can only be arranged in three distinct ways



each of the three protons of the CH_3 group can couple with their neighbor in just three ways; the result is the splitting of its spectral line into a triplet.

The extent of the band splitting is expressed in terms of a coupling constant, J, that, unlike the chemical shift, is independent of the magnitude of the external B-field. Typically, J is of the order of a few Hertz.

6.3.6 Applications of Nuclear Magnetic Resonance

Since its initial development, nuclear magnetic resonance has become a standard research and diagnostic tool. Initially, its major application was in the study of molecular structures.

By measuring chemical shifts and examining the breadth and fine structure of the spectral lines, a vast quantity of detailed and accurate information could be obtained about the shape and environment of molecules.

Important information concerning the structure of biological membranes and the proportions of phospholipid vesicles and detergent and bile-salt micelles has been obtained from the NMR spectrum of the $_1$H, ^{2}D, ^{13}C and ^{31}P nuclei incorporated in these species.

Well defined narrow spectral lines were found in the spectra of small vesicles and micelles (radii up to 20nm).

However, the lines widened and even disappeared in large multi-layered vesicles (radii of ~l μm).

Using the NMR spectrum of ^{31}P, phase transitions from bilayers to other configurations have been identified in membranes.

Similarly, increases in the viscosity of phospholipid membranes caused by cations such as Ca++, Mg++ and Na+ have been detected.

More recently, techniques have been developed by which images of the structure of individual cells, tissues and even whole organisms can be obtained using nuclear magnetic resonance.

The technique is called NMR Imaging or Magnetic Resonance Imaging.

The process hinges on two factors:

1. The dependence of the resonant (Larmor) frequency on the strength of the magnetic field;

2. The dependence of the relaxation time on the environment of the nucleus.

If a specimen containing nuclear magnetic moments is placed in a non-uniform magnetic field (a magnetic field with a spatial gradient), the resonant frequency exhibited by each nucleus will depend on its position in the field. If the contours (the map) of the magnetic field are known, the observed frequency of given nuclei in the specimen, say the protons, will specify their spatial co-ordinates. In this way, the nuclei on one side of the specimen can be distinguished from those on the other side (Fig 6.23). By changing the contours or the orientation of the field, other sets of co-ordinates can be obtained. Combining the various sets of co-ordinates produces a two-dimensional image of the distribution of the nuclei in the specimen.

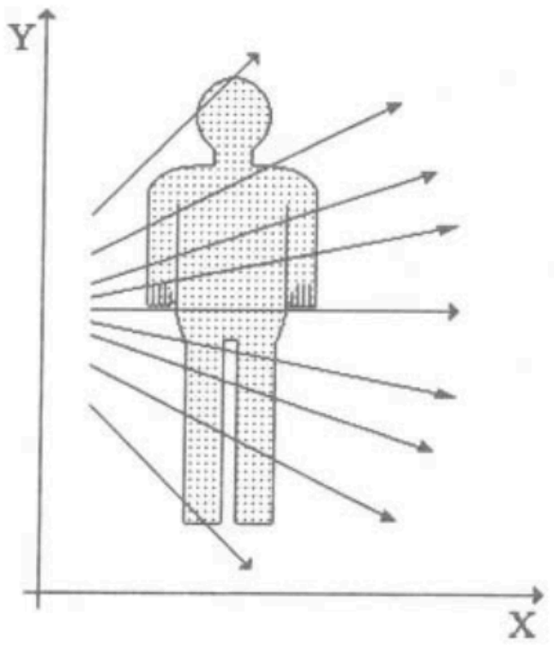


Fig 6.23 A non-uniform magnetic field. The field at each point is a function of the point's coordinates.

Protons in tissues such as subcutaneous fat and bone marrow have shorter relaxation times than those in fluids such as the cerebrospinal and ocular fluids.

These differences are revealed by applying the oscillating magnetic field, B', in carefully timed pulses of a fairly broad band of frequencies.

If the time period, τ, between the pulses is less than the relaxation time, T, of a particular nucleus in the sample, the nucleus will not respond to subsequent pulses.

The first pulse elevates the nucleus into the relatively stable higher state and as long as it remains in this elevated state, it will not respond to the later pulses.

Only those nuclei whose environment does not hinder their relaxation into the ground state, i.e., those nuclei whose relaxation time, T, is less than τ, will respond to the later pulses.

These differences between the nuclei are detected electronically and add contrast and detail to the two-dimensional image of the nuclei in the specimen.

By taking a sequence of images as the sample traverses the magnetic field, 'whole-body' scans can be obtained.

In order to obtain the strong (>1T) and spatially large magnetic fields required for human whole-body imaging, superconducting magnets are used.

The use of NMR. imaging in medicine has two significant advantages over X-ray imaging (tomography):

* There is no evidence that magnetic fields have a detrimental effect on the human body, whereas the damage caused by ionising radiation such as X-rays is well documented.
* X-ray images are essentially just pictures of the electron density in the specimen; NMR images are sensitive to a wide range of different chemical and physical conditions in the tissue and so can provide more information.

Part 6.4 The Conduction of Electricity Through Solids

No single development of modern physics has touched the lives of people everywhere more than the invention of the transistor.

Requiring less energy to process a single unit of information -a bit- than do brain neurons, transistor based devices have reduced the cost of computing and of information to an unprecedented level.

By revolutionizing electronics, they have put the whole world in the lap of every human being on earth.

How this was done is the central theme of this section.

6.4.1 The Electrical Conductivity of Solids

The electrical conductivity of a solid material is generally specified in terms of the following three parameters:

1. the material's *resistivity*, ρ,
2. the *temperature coefficient*, α, of its resistance;
3. the *number of charge carriers*, N, in each unit volume of the material.

A material's resistivity indicates the opposition it presents to the passage of an electric current.

At room temperature, certain materials, such as diamond and PVC, do not conduct electricity to any measurable degree.

Such materials are called insulators; their resistivity is exceedingly high: $\rho_{diamond} \approx 10^{16} \Omega$-m.

On the other hand, metals like copper and silver present very little opposition to the passage of an electric current.

Such materials are called conductors.

Copper and silver are excellent conductors of electricity; their resistivity is very low: $\rho_{copper} \approx 10^{-8} \Omega$-m.

The resistivity of a material depends inter a/ia on its temperature.

Typically, the resistivity of metals increases with temperature, i.e., the temperature coefficient of their resistance is positive: $\alpha_{copper} \approx +4 \text{x } 10^{-3} K^{-1}$.

Metals also contain large numbers of charge carriers: $N_{copper} \approx 8.5 \text{x} 10^{28} m^{-3}$.

These are the particles whose flow through the material constitutes the electric current.

The charge carriers in metals are generally electrons.

A third group of materials, called the semiconductors, have properties between those of the insulators and metal conductors; silicon is a typical semiconductor.

As the data in Table 6.2 show, the resistivity of the semiconductor silicon, lies between that of the insulator diamond and that of the conductor copper and it contains fewer charge carriers per unit volume than copper.

Table 6.2: The Electrical Conductivity Parameters of Diamond, Copper and Silicon

	Diamond	Copper	Silicon
ρ - Resistivity ($\Omega \cdot m$)	10^{16}	$1.6 \cdot 10^{-8}$	$3 \cdot 10^{3}$
α - Temperature Coefficient of Resistance (K^{-1})	N/A	$+4 \cdot 10^{-3}$	$-70 \cdot 10^{-3}$
N - Number of Charge Carriers per Unit Volume (m^{-3})	Negligible	$8.5 \cdot 10^{28}$	10^{16}

Like so much in modem physics, the study of semiconductors originated with Faraday.

As long ago as 1833, Faraday discovered that in contrast to regular metallic conductors, the resistance of certain poor electrical conductors (he was working with silver sulphide) falls with rising temperature.

That the conductivity of semiconductors was essentially different from that of regular metallic conductors became evident from measurements of the *Hall effect*.

Hall had discovered that if a strip of conducting material is placed in a magnetic field whose direction is perpendicular to its length and an electric current flows along the strip, a potential difference is set up between the edges of the strip, i.e., perpendicular to the direction of the current.

The direction of this Hall potential depends on the sign of the charged particles that carry the current flowing through it.

Measurements of the Hall potential showed that whereas the charge carriers in regular metallic are invariably negatively charged, those in semiconductors may be negatively or positively charged.

Moreover, whereas the number of these charge carriers is virtually independent of the temperature in metallic conductors, their number increases greatly with temperature in semiconductors.

By the end of the 19th century, three other important properties of these materials had been discovered, all of which have since found numerous practical applications:

1. **The *photovoltaic effect*: the production of an electromotive force (an e.m.f.) when a junction of dissimilar materials, one of which is a semiconductor, is illuminated by visible or near ultraviolet light. The junction may be an electrolytic cell or simply a thin mesh of the semiconductor material overlaid on a metal substrate. *Solar cells* are a development of this effect.**
2. **The rectifying property of a metal-semiconductor junction: the fact that such junctions do not obey Ohm's law. Their resistance—the ratio between the voltage across them and the current passing through them—is not a constant but depends on the sign as well as the magnitude of the voltage.**
3. ***Photoconductivity*: the increased conductivity of a semiconductor when it is illuminated by visible or near ultraviolet light. It is the basis of the dry copying process known as *xerography*, by which powdered ink—'toner'—is transferred from an optically induced image on a negatively charged transfer device to a piece of positively charged paper. The transfer device is a rotating drum coated with selenium, which retains electrostatic charge almost indefinitely in the absence of light, but loses it at once wherever light falls upon it. The image is 'etched' onto the charged drum by projecting light onto its surface. Wherever the light strikes, charge is lost; elsewhere, it is retained. Toner, is next sprayed onto the drum and sticks to the charged areas. The drum finally passes over a positively charged sheet of paper onto which the toner is transferred and fixed by heat treatment.**

6.4.2. The Electron Gas

A short time after the discovery of the electron, Lorentz and Drude proposed a simple classical model to explain the electrical conductivity of metals.

The outstanding feature of this model was its hypothesis that a 'gas' of free electrons pervades the interstices of metallic crystals.

It was postulated that the lattice energy released as a metal crystallizes, frees the valence electrons from its atoms, leaving the atoms positively charged.

Thus the model postulated that a metal is composed of a lattice of positive metal ions whose interstices are filled by the gas of free valence electrons.

The excellent electrical conductivity of metals was credited to the relative ease with which this gas of conduction electrons, as they were subsequently called, could move through the lattice under the influence of an electric field.

The electron gas was regarded at the time as being essentially no different from any other gas such as hydrogen or oxygen.

Thus, the average velocity, v, due to their random thermal motion at an absolute temperature T, of the electrons comprising the gas was calculated from the classical relationship

$$\frac{m_e v^2}{2} = \frac{3}{2} kT \tag{6.44}$$

which gives a value at room temperature of $v \approx 10^5 m/s$.

Placing a metal in an electric field causes the conduction electrons to flow.

Because of their negative charge, the electrons are accelerated in the opposite direction to that of the field.

As their velocity increases, they acquire kinetic energy and linear momentum in the direction of their flow.

However, in the classical model, the progress of the conduction electrons through the lattice is slowed by their collisions with the metal ions, in which they lose both energy and momentum to the lattice.

When the average rate at which the conduction electrons acquire energy and momentum from the electric field is balanced by the average rate at which they lose energy and momentum to the lattice ions, a steady electric current flows and heat is generated in the metal at a constant rate.

It can be shown, mathematically, that the classical model predicts both Ohm's Law, i.e., the direct proportionality of electric current to potential difference, and the positive value of the temperature coefficient of a metal conductor's resistance.

However, its quantitative predictions of the resistivity of metals and of its temperature dependence are not borne out by experiment.

For instance, whereas the model predicts that a metal's resistivity should be a function of the square root of its absolute temperature, i.e., $\rho \propto \sqrt{T}$, at low temperatures (30 < K <120) resistivity is actually found to vary approximately linearly with temperature, that is $\rho \propto \mathrm{T}$.

The classical model takes no account of the wave nature or spin of the electron; this is a fundamental flaw.

Because electrons are fermions and are subject to the exclusion principle, the electron gas cannot be compared to an ordinary gas; electrons obey a different statistics from normal gas molecules.

Furthermore, the intuitive idea that the conduction electrons must collide with the metal ions as they flow through the lattice is untenable in wave mechanics.

Matter waves propagate unhindered through regular structures and so a regular lattice of positive ions would not oppose the flow of the conduction electrons at all; they would only be scattered from any deformations there may be in the lattice.

And thirdly, electrons can only accept energy from the applied field in discrete packets, rising each time from one energy level to a higher one.

If there are no higher levels both vacant and accessible, the electron cannot accept the energy and a current cannot flow.

The use of quantum mechanics in the study of the electrical conductivity of solids has been one of its most fruitful applications.

A full account of this achievement is beyond the scope of this class.

The following descriptive account is intended just to outline its underlying principles and some of its more outstanding results.

The Relaxation Time of Conduction Electrons

Assuming they are repeatedly accelerated by the electric field and halted by collisions with the lattice, we can find the average time, τ, (the *relaxation time*) the conduction electrons in copper are actually in motion between collisions.

Consider a metal conductor of length L and uniform cross-section S, between whose ends a uniform electric field, E, is applied (Fig 6.24).

Ohm's Law relates the potential difference, V = EL, between the ends of the conductor to the current, I, flowing along it.

$$V = EL = IR \tag{6.45}$$

R is the electrical resistance of the conductor.

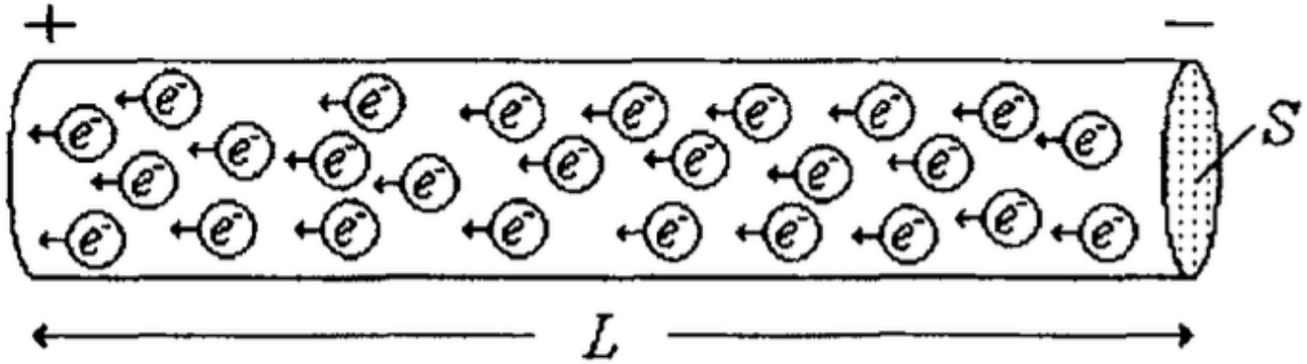


Fig 6.24 The drift of the conduction electrons along a uniform metal conductor between whose ends an electrical potential is applied.

In terms of the drift velocity, v_d, of its conduction electrons along the conductor, the electric current flowing through a metal conductor of cross-section S is given by:

$$I = NSev_d \tag{6.46}$$

where N is the number of conduction electrons in each unit volume of the conductor.

In terms of the resistivity, ρ; of the conducting material, the resistance of a conductor of length L and cross-section S is

$$R = {}^{L\rho}/_{S} \tag{6.47}$$

Combining equations (6.45), (6.46) and (6.47) gives

$$\rho = \frac{E}{Nev_d} \tag{6.48}$$

The electric field exerts a force of magnitude F= eE on the electrons in the conductor. Applying Newton's second law of motion, $a = F/m_e$, to the conduction electrons gives:

$$a = \frac{eE}{m_e} \tag{6.49}$$

Hence, the drift velocity, $v_d = a\tau$ of the electrons is given by:

$$v_d = \frac{eE}{2m_e} \cdot \tau \tag{6.50}$$

Substitution in equation (6.48) gives:

$$\tau = \frac{2m_e}{Ne^2\rho} \tag{6.51}$$

The relaxation time of copper obtained from the values of the parameters in Table 6.2 is:

$$\tau = \frac{2 \cdot 9.11 \cdot 10^{-31}}{8.5 \cdot 10^{28} \cdot \left(1.6 \cdot 10^{-19}\right)^2 \cdot 1.6 \cdot 10^{-8}} = 5 \cdot 10^{-14}\,\mathrm{s} \tag{6.52}$$

6.4.3 Energy Levels in Solids - Band Theory

Free atoms, such as those in sodium vapor, can be regarded as isolated systems whose energy levels are determined by a single set of quantum numbers.

However, in an aggregate of atoms such as a liquid or solid, where the atoms are close to one another, the energy levels of each atom are affected by the presence of the other atoms.

In terms of quantum mechanics, the wave functions of the atoms overlap.

The wave function describing the aggregate is obtained from the appropriate superposition of the wave functions of the individual atoms.

The effect that the approach of another atom has on a system's wave function is most simply shown by considering the convergence of two ground state hydrogen atoms.

So long as the atoms are widely separated, their wave functions are autonomous and their electrons each have the same energy; both occupy the ls state.

However, as the atoms converge, their orbitals (spatial wave functions) interfere, either constructively to give the symmetric orbital $\phi_S = \phi_1 + \phi_2$ (covalent bond formation that produces the H_2 molecule) or destructively to give the anti-symmetric orbital $\phi_A = \phi_1 - \phi_2$ (the anti-bonding orbital).

The atomic electrons can now occupy either the anti-symmetric or the symmetric orbital.

The symmetric orbital corresponds to a lower energy than the anti-symmetric orbital.

When two hydrogen atoms converge the number of electronic energy levels is doubled.

It is tripled by the convergence of three atoms, quadrupled by the convergence of four atoms, and so on.

The cluster of energy levels that corresponds to each atomic energy level is called an energy band (Fig 6.25).

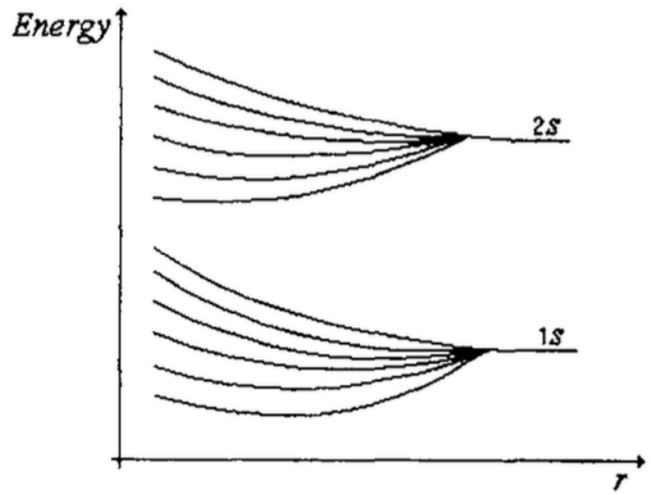


Fig 6.25 The energy levels of six hydrogen atoms arranged in a straight line as a function of the distance, r, between them. Each atomic level is split into a band comprising six sub-levels.

In general, the number of sub-levels in each band is equal to the number of atoms in the system.

Accordingly, the number of electrons required to fill a band in a system comprising N atoms is 2N(2l + 1), where l is the orbital quantum number of the corresponding atomic energy level.

The exclusion principle permits the deployment of no more than two electrons in each sub-level.

Thus, 2N electrons are needed to fill a band corresponding to an atomic s state and 6N will be needed for one corresponding to a p state.

The atoms in almost every crystalline solid, whether metallic or non-metallic, are so close to one another that their valence electrons constitute a single distinct system that pertains to the entire crystal.

As a result, each atomic energy level is replaced by an energy band comprising a myriad of discrete sub-levels.

The typical width of an energy band is a few electron-volts.

Since, the average separation between the sub-levels comprising a band is so very small, ~10^{-23}eV, each may in practice be regarded as a being a continuum of energy values.

The width of the energy bands depends on the extent to which the orbitals overlap.

In aggregates of multi-electron atoms, it is the orbitals of the outermost electrons-the valence electrons-that overlap the most.

Consequently, the broadest energy level bands are those associated with these electrons.

Commonly, these outermost bands are so broad that they overlap.

For example, whereas the inner 2p band in metallic sodium is unmoved under usual conditions, the bands corresponding to the 3s and 3p atomic levels overlap one another.

The bands corresponding to the closed inner atomic shells are narrower and are usually full.

The electrons in these inner bands are unable to change their state; they are 'frozen' in them.

Figure 6.26 shows typical band configurations in systems in which the bands do not overlap.

The conduction bands are so called because it is the presence or otherwise of electrons in this band that determines the electrical conductivity of the solid.

The gaps between the bands indicate forbidden energy values, like the gaps between atomic energy levels; they are often referred to as forbidden bands.

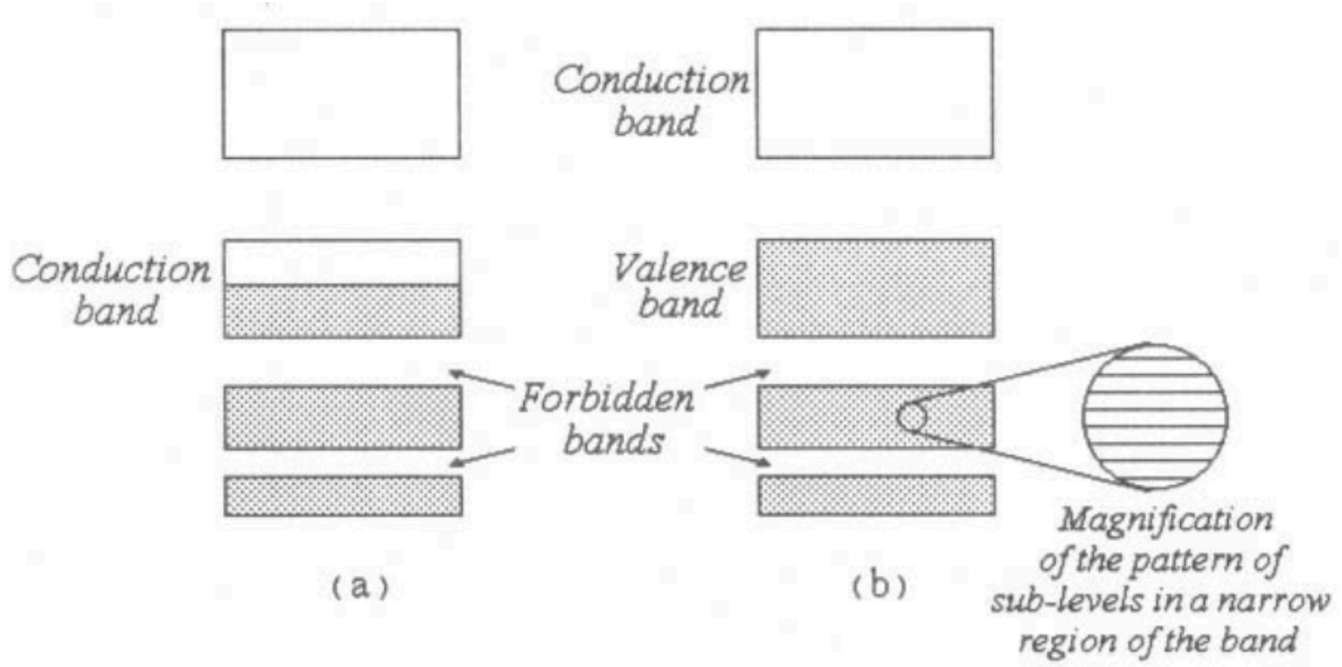


Fig 6.26 Typical energy band diagrams of two different solids. Each band comprises a myriad of very closely spaced sub-levels. The lower energy bands are narrower since the orbitals of these inner electrons overlap less. When, as in (a), the highest populated band is only partially filled, it is called the conduction band. When the highest populated band is full, as in (b), it is called the valence band and the empty band above it is the conduction band.

6.4.4 Insulators

TheDiamond is a pure transparent crystalline form of carbon; it is a typical solid insulator.

The electronic configuration of the carbon atom is $1s^2 2s^2 2p^2$. At the inter-atomic distances found in diamond, the 2s and 2p energy levels form two bands each of which can accommodate up to 4N electrons.

The four outermost electrons of each atom are normally found in the lower valence band; this leaves the higher conduction band empty.

There is a significant energy difference of ~ 6eV between the two bands (Fig 6.27).

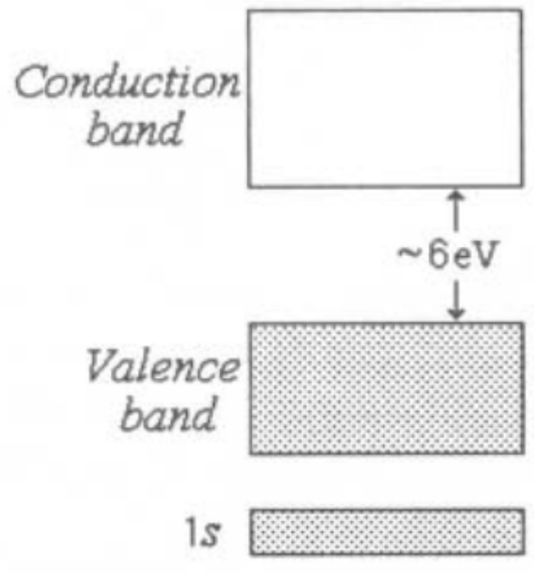


Fig 6.27 Energy band diagram of diamond. There are no vacant states in the valence band. The excitation of an electron from the valence band to the conduction band requires at least 6eV.

There are no vacant states in the diamond's valence band.

On the other hand, there are normally no electrons in the conduction band.

This situation sets a minimum on the amount of energy an electron in diamond can absorb.

The only transition open to an electron is to rise from the valence band to the conduction band, but this requires at least 6eV of energy.

At ordinary temperatures, the probability that an electron will accumulate this amount of energy from the thermal energy of the lattice is in effect nil.

The average thermal energy of a lattice particle is equal to kT, which at 300K is only $8.62x10^{-5}x300 = 0.025eV$.

Nor, in practice, can the electrons be physically displaced.

Assuming an electron might move a distance of about $l0^{-8}m$ through a diamond before being halted by a collision with a lattice particle, the electric field required for it to acquire 6eV during the move would be $6x10^{8}V/m$.

Such a field is 10^{10} times stronger than that needed to establish an electric current in copper.

This explains why diamond is such an excellent insulator.

The large energy gap between its valence and conduction bands also explains the transparency of diamond.

The photons in visible light have energies between 1.8 and 3.1 electron-volts.

However, the minimum amount of energy an electron in diamond can accept is 6eV. Consequently, the photons of visible light pass through diamond unabsorbed.

6.4.5 Metallic Conductors

For the purposes of the descriptive account given here, we shall assume that the electrons moving through a metal conductor have zero potential energy, i.e., they possess only kinetic energy.

In fact, electrons experience a periodic potential as they move through a metal lattice; the form of this potential being determined by the lattice structure.

However, for the purposes of the elementary treatment given here, the periodic nature of the electric potential in metal conductors will be ignored.

Sodium is a typical metal conductor.

The electronic configuration of the sodium atom is $1s^2 2s^2 2p^6 3s^1$.

In the metal solid the bands corresponding to the ls, 2s and 2p energy levels are fully occupied, but the band corresponding to the 3s level is only half-full.

Thus the 3s band is the conduction band (Fig 6.28).

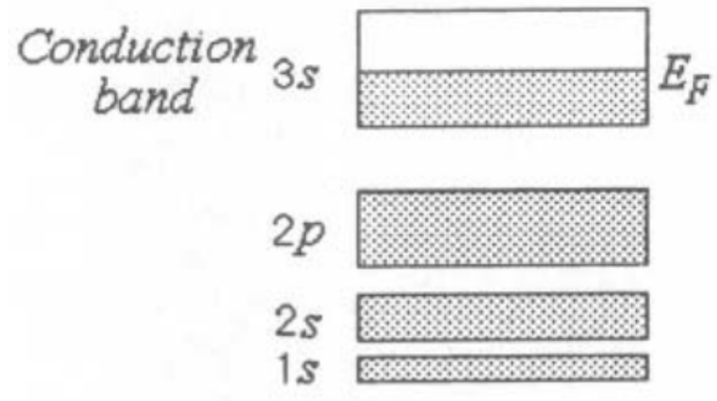


Fig 6.28 Energy band diagram of sodium metal at 0K. The electrons in the conduction band have kinetic energies from zero up to the Fermi energy, E_p.

The appropriate function for the probability, $f(E_i)_{Electron}$, of an electron occupying the E_i energy level in any system at an absolute temperature T is that given by the Fermi-Dirac statistics, namely,

$$f(E_i)_{electron} = \left(Ce^{E_i/kT} + 1 \right)^{-1} \tag{6.53}$$

In the case of the electron gas in a metal conductor the parameter C is the exponential function

$$C = e^{-E_F/kT} \tag{6.54}$$

E_F is called the Fermi energy of the metal.

Thus, the probability, $f(E_i)_{Electron}$, of an electron occupying the E_i energy level in the conduction band in a metal conductor at an absolute temperature T is given by

$$f(E_i)_{electron} = \left(e^{(E_i - E_F)/kT} + 1 \right)^{-1} \tag{6.55}$$

It follows, that at a temperature of 0K (absolute zero), the probability of finding a conduction electron in an energy level $E_i < E_F$ is:

$$f(E_i)_{electron} = \left(e^{(E_i - E_F)/kT} + 1 \right)^{-1} = \left(e^{-\infty} + 1 \right)^{-1} = 1 \tag{6.56}$$

i.e., all such energy levels will be occupied.

On the other hand, the probability of finding a conduction electron in an energy level $E_i > E_F$ is:

$$f(E_i)_{Electron} = \left(e^{(E_i - E_F)/kT} + 1 \right)^{-1} = \left(e^{\infty} + 1 \right)^{-1} = 0 \tag{6.57}$$

i.e., all such energy levels will be empty.

Thus, at absolute zero, all the energy levels up to the Fermi energy are occupied by electrons, whereas all those above it are empty.

From the values of the Fermi energy given in Table 6.3, it is clear that even at absolute zero, many of the electrons in a metal conductor possess a kinetic energy of several electron volts.

Classical gas molecules must be heated to temperatures of tens of thousands of degrees for them to possess kinetic energies of this magnitude.

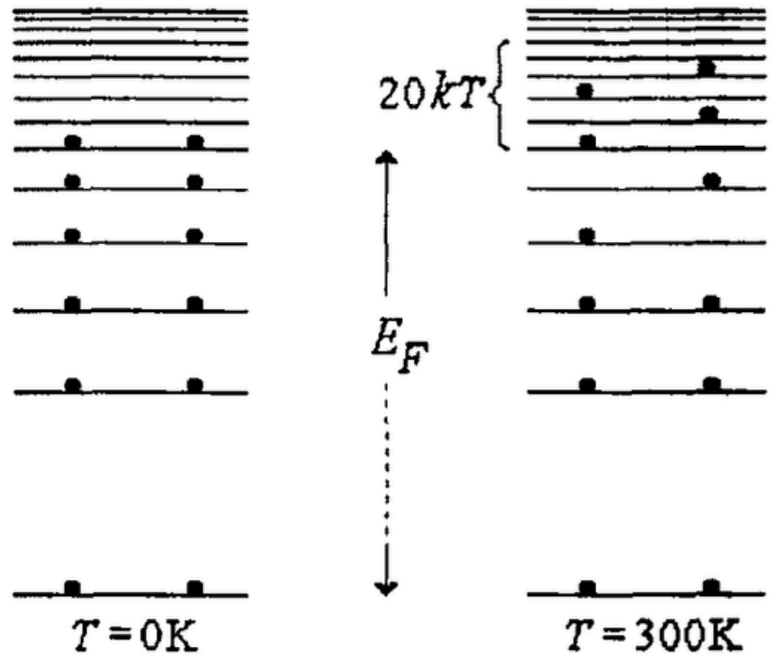
Table 6.3: Typical Fermi Energies

Metal		Fermi Energy (eV)
Sodium	Na	3.12
Caesium	Cs	1.53
Copper	Cu	7.04
Zinc	Zn	11.0
Silver	Ag	5.51

The additional thermal energy possessed by the electrons at higher temperatures -0 .025eV at room temperature- is almost insignificant compared to the Fermi energies.

Consequently, the electronic configuration of a metal conductor changes only slightly as its temperature rises (Fig 6.29).

Fig 6.29 The energy levels of the conduction electrons in a metal at absolute zero, T = 9K, and room temperature, T = 300K. The excited levels occupied by electrons at room temperature cover a range of about 20 kT above the Fermi energy.



Little of the heat energy absorbed by a metal as it warms is taken up by the electron gas; almost all of it goes to increase the thermal vibrations of the lattice.

The key feature of the electronic structure of metal conductors is the abundance of empty electronic states in the conduction band just above the Fermi energy.

As a result of these, the conduction electrons occupying energy levels close to the Fermi level can absorb a whole range of energies, from a just a fraction of an electron volt to several electron volts, thereupon rising to one of the many vacant levels.

This accounts for (i) the opacity of metals and (ii) their excellent electrical conductivity.

> **(i) The photons in visible light have energies between 1.8 and 3.1 electron-volts. These are readily absorbed by the conduction electrons occupying energy levels close to the Fermi level and so metals are opaque.**
>
> **(ii) In the presence of an electric field, even a very weak field, conduction electrons occupying energy levels close to the Fermi level can easily 'jump' into the vacant higher levels, thereby absorbing energy from the field. Under the influence of the field, the electrons flow through the metal lattice producing an electric current.**

The mechanism by which a steady electric current arises and heat is continuously generated in the metal is similar in the quantum-mechanical model to that sketched by the classical model.

Both models posit collisions between the electrons comprising the electric current and the particles constituting the lattice, such that the rate at which the conduction electrons absorb energy from the applied field is balanced by that at which they lose energy to the lattice.

However, the lattice particles involved in the transfer of the enengy differ in each model.

Whereas in the classical model the particles with which the conduction electrons collide are the metal ions that constitute the lattice, in the quantum-mechanical model they are the defects in the lattice.

The metal ions themselves do not impede the propagation of the electron waves; matter waves pass unhindered through a regular lattice.

Two different types of lattice defects impede the passage of the electrons:

1. **Defects in the structure and arrangement of the lattice itself. For example, atoms that have been displaced from their position in the structure or impurity atoms that have been incorporated into the lattice. These defects are independent of temperature.**
2. **Thermal vibrations of the lattice that result in random variations in the spacing of the lattice ions. These defects are temperature dependent; the higher the temperature, the more vibrational defects there are.**

The positive temperature coefficient of resistance typical of metal conductors is accounted for by the temperature dependent lattice defects.

Raising the conductor's temperature increases the number of lattice defects.

As a result, the conducting electrons are scattered to a greater degree and a weaker current is produced in the metal conductor by a given electric field, i.e., the conductor's resistance increases.

6.4.6 Superconductivity

In 1911, Kammerlingh Onnes discovered that at temperatures below 4.15K the metal mercury exhibits zero resistivity.

Other superconducting metals and the critical temperatures, T_C, below which they exhibit this remarkable property are listed in Table 6.4.

Table 6.4: Metal Superconductors

Metal		T_C(K)
Aluminum	Al	1.19
Tin	Sn	3.7
Lead	Pb	7.18
Zinc	Zn	0.55
Niobium	Nb	9.46

Note that the best normal conductors such as silver and copper do not exhibit superconductivity, suggesting that normal conductivity and superconductivity are fundamentally different phenomena.

The discovery by Bednorz and Muller in 1986, that certain ceramic materials, which are insulators at room temperature, exhibit superconductivity when cooled to temperatures around 100K, reinforces this impression.

These materials are remarkable as they exhibit superconductivity at much higher temperatures, above the boiling point of nitrogen (77K), than do metal superconductors.

Superconductors are not only perfect conductors, they are also perfect diamagnets, i.e., the magnetic field within a superconductor is always zero.

The applied magnetic field is neutralized inside the superconductor by eddy currents induced in its surface.

The expulsion of the magnetic field from a superconductor situated in a magnetic field is called the Meissner effect.

As a result of the exclusion of magnetic flux from the superconductor, a bar magnet placed over the superconductor levitates above its surface (Fig 6.30).

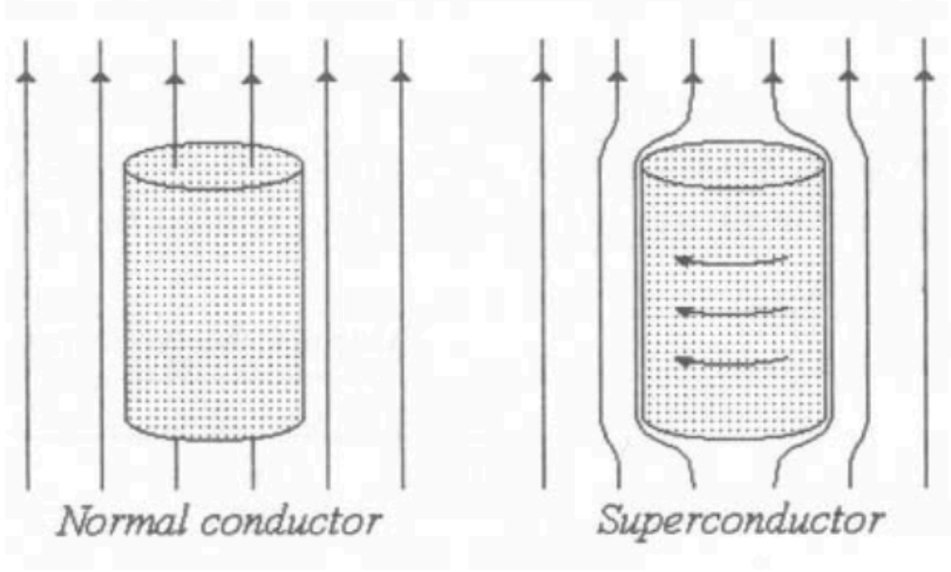


Fig 6.30 The Meissner effect. As a metal situated in a magnetic field changes from a normal conductor into a superconductor eddy currents are set up in its surface which expel the magnetic flux that was passing through it.

A theory that elucidates the mechanism underlying superconductivity was only spelled out some sixty years after its initial discovery.

The theory, known as the Bardeen-Cooper-Schrieffer (BCS) theory, postulates that the charge carriers in superconductivity are not single electrons but pairs of electrons.

At the very low temperatures at which superconductivity is exhibited, a lattice mediated attractive interaction acts between the conduction electrons as a result of which they are paired.

As they move through a conductor, the negatively charged electrons cause slight distortions in the lattice of positive ions, distortions that are so slight that they only assume significance when the thermal vibrations of the lattice become insignificant, i.e., at very low temperatures.

At these low temperatures, the net effect of the local intensifications of positive charge associated with these lattice distortions is to couple the electrons into pairs called Cooper pairs.

The involvement of the lattice in superconductivity has been demonstrated by the effect on the transition temperature, Tc, of the presence of different isotopes in the lattice.

If one of the electrons comprising a Cooper pair has spin -1/2 and linear momentum p, its partner will have spin +1/2 and linear momentum -p.

In classical particle terms this is equivalent to saying that they are moving in opposite directions, a concept that defies intuition; how can particles moving away from each other maintain a coupling interaction?

However, in quantum mechanical terms the particles can continue to interact as long as their matter waves overlap.

The implication is that the wave function of a Cooper pair extends over a large volume, so large a volume that it overlaps the wave functions of as many as a million other Cooper pairs.

Hence the Cooper pairs in a superconductor can be regarded as occupying a collective state that extends over the whole volume of the superconductor; all the Cooper pairs occupying this state cooperate with one another.

This configuration of the conduction electrons is reminiscent of the behavior of bosons and has been interpreted as indicating that the superconducting state should be regarded as a distinct state of matter, essentially different from the familiar solid, liquid, gas and plasma states.

According to the BCS theory, the conduction electrons in a superconductor can be regarded as having condensed into a single 'macromolecule' that pervades the entire volume of the system and can flow through it as a whole.

An energy gap of order 10^{-3}eV separates the electrons in the 'macromolecule' from the first excited state and thus ensures its stability.

This gap represents the energy required to split a Cooper pair and it accounts for the essential properties of superconductors.

1. The 'macromolecule' of conduction electrons moves unhindered—without resistance—through the superconductor because the energy of its friction with the lattice irregularities is insufficient to enable electrons to separate from their Cooper partners and rise into the excited state.
2. Under the influence of the external magnetic field, the motion of the conduction electrons in the superconductor is accelerated such that eddy currents are set up whose secondary magnetic fields oppose the external field. This is the familiar phenomenon of diamagnetism. However, so long as the external field is not too strong, the acceleration of the electrons is not accompanied by a transfer to a higher energy level; the 'macromolecule' of Cooper pairs remains intact and so there is nothing to prevent eddy currents being induced that totally neutralize the magnetic field inside the superconductor.

6.4.7 Semiconductors

The configuration of the energy bands in pure (intrinsic} semiconductors is similar to that in insulators except that the gap between the valence and conduction band is smaller.

For example, the energy gap, Eg, in silicon is just l. leV (Fig 6.3la).

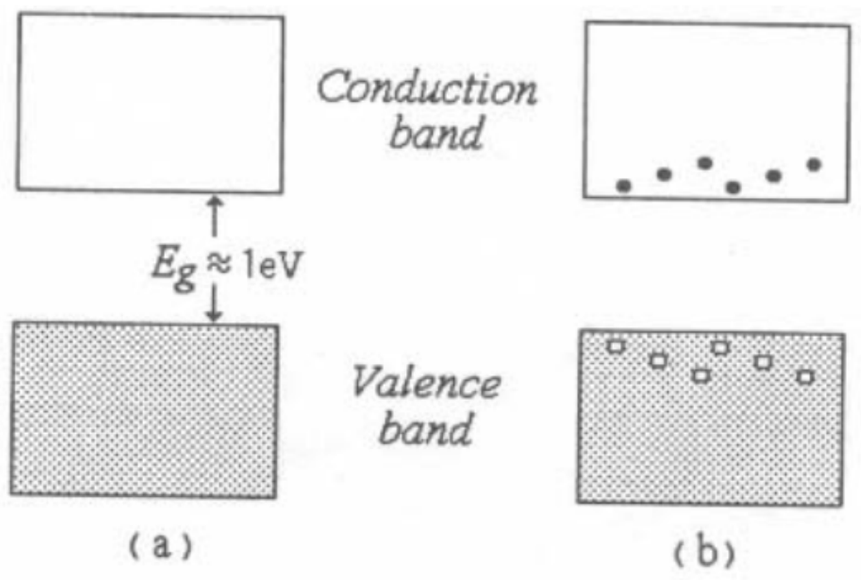


Fig 6.31 (a) The energy band diagram of an intrinsic semiconductor at 0K; the valence band is full and the conduction band is empty. (b) At higher temperatures a few electrons (represented by the black dots) acquire sufficient thermal energy to rise from the valence to the conduction band leaving an equal number of 'holes' (represented by the empty squares) in the valence band

At low temperatures, silicon is as poor a conductor of electricity as diamond.

However, as the temperature rises, the probability of a few electrons acquiring sufficient thermal energy from the lattice to cross the energy gap separating the valence from the conduction band increases (Fig 6.31b).

At a temperature of 450K the average population of the conduction band is a million times greater than at 250K.

Thus, the temperature coefficient of resistance, a, of semiconductors has a negative value, i.e., their resistance falls with rising temperature.

An electron that happens into the semiconductor's conduction band leaves a vacancy or positive 'hole' in the valence band.

This hole can subsequently be filled by an electron that drops back into the valence band from the conduction band.

Thus, positive holes are continuously and randomly created and annihilated in the semiconductor lattice.

The holes are treated like electrons except that their charge is positive and their energy in energy level diagrams is measured downwards.

The charge carriers in intrinsic semiconductors are both the electrons in the conduction band and the holes in the valence band.

As the conduction electrons drift through the lattice in one direction, the holes they left in the valence band drift through it in the opposite direction.

The maximum current an intrinsic semiconductor can carry -its saturation current- is determined by the number of conduction electrons and positive holes it contains and their respective mobilities.

Until the saturation current is reached, intrinsic semi-conductors generally obey Ohm's law.

The presence of even minute amounts (one part in a million) of certain impurities has a dramatic effect on the conductivity of intrinsic (pure) semiconductors such as silicon.

Without causing a distortion to the lattice, these trace impurities can increase the number of charge carriers in the material by a factor of a million or even more.

These doped materials are called extrinsic semiconductors.

Two types of materials are usually used as the impurities, Group III elements and Group V elements.

* **Semiconductors doped with traces of Group V pentavalent elements such as phosphorus, arsenic and antimony are called n-type semiconductors.**
* **Semiconductors doped with traces of Group III trivalent elements such as aluminium, indium and gallium are called p-type semiconductors.**

Two-dimensional schematic representations of the lattices of intrinsic (pure), n-type and p-type semiconducting materials are shown in Fig 6.32.

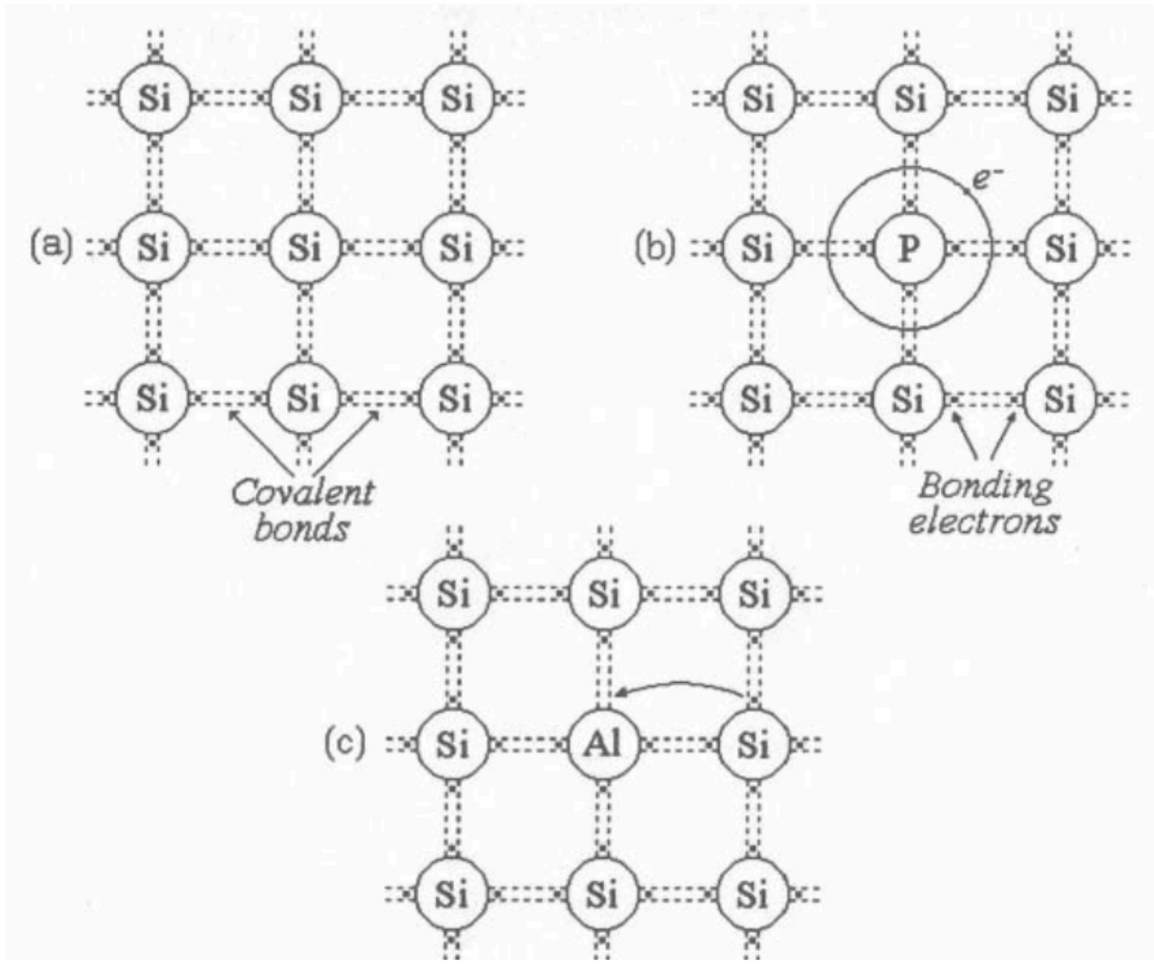


Fig 6.32 Two-dimensional schematic representations of semiconductor lattices. (a) Pure silicon; each quadrivalent silicon atom is bonded covalently to its four closest neighbors. (b) An n-type semiconductor in which a pentavalent phosphorus atom replaces one of the silicon atoms; the surplus fifth valence electron is only loosely held to the phosphorus atom. (c) A p-type semiconductor in which a trivalent aluminum atom replaces one of the silicon atoms; the missing electron is equivalent to a positive hole which effectively moves randomly through the lattice as it is repeatedly filled by an electron only to reappear at the site previously occupied by the electron.

In n-type semiconductors, the fifth valence electron in the Group V atoms is not required for bonding the atoms in the lattice.

This non-bonding electron is only loosely bound to its parent atom (Fig 6.32b).

The energy, E_D, it requires to break loose from its parent atom is typically -0.05eV.

Consequently, at room temperature, the electron can easily escape from the atom and wander freely through the lattice; the necessary energy is obtained from the thermal vibrations of the lattice.

In terms of energy levels, the surplus electrons in n-type semiconductors occupy levels that are just below the semiconductor's conduction band, i.e., $E_D << E_g$.

These electrons can easily rise from these donor levels into the semiconductor's conduction band, thereby enhancing its population and the conductivity of the material (Fig 6.33).

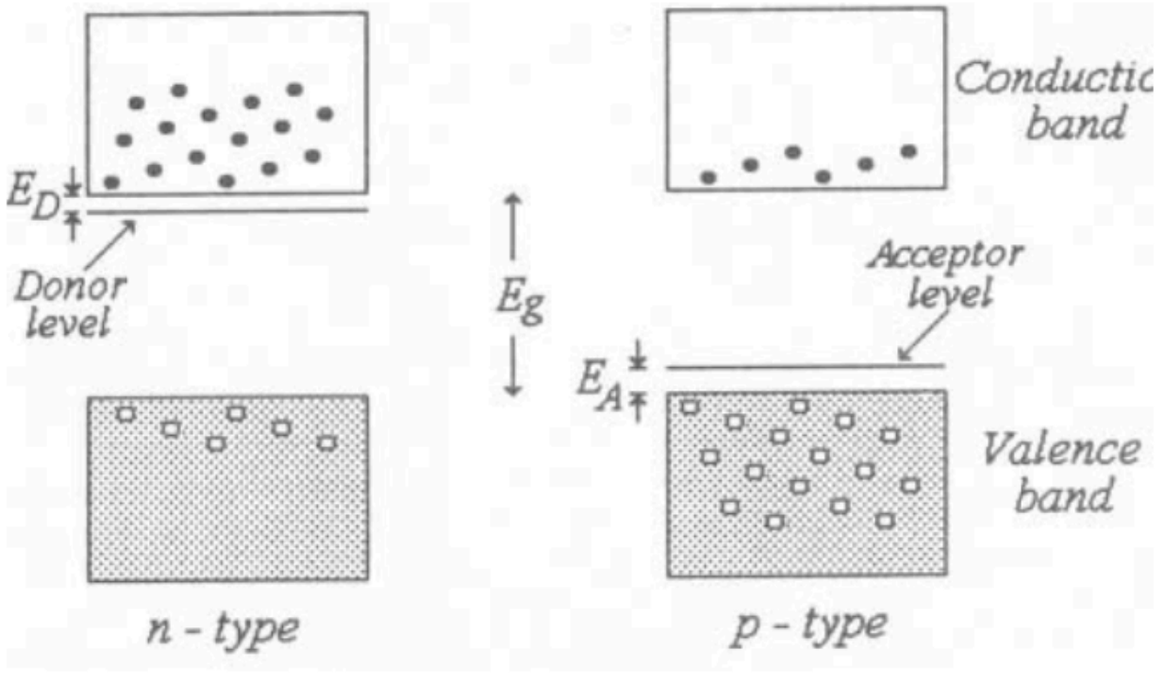


Fig 6.33 The energy bands and energy levels in n-type and p-type semiconductors. In the n-type, most of the electrons in the conduction band come from the donor level and just a few fror the valence level. In the p-type, the valence band contains man holes that are the result of electrons being promoted to the acceptor level and just a few that are the result of electrons promoted to the conduction level. The electrons remaining in the donor level of the n-type material as well as those that have risen to the acceptor level in the p-type are not shown in the diagram as they make no direct contribution to the materials' electrical conductivity.

This. however, does not increase the number of labile positive holes in the valence band because the positive charge that balances the negative charge of the surplus electrons is 'locked' in the nuclei of the donor Group V atoms.

Thus, it is the electrons in the conduction band, that constitute the principle (majority) charge carriers in n-type semiconductors.

The holes in the valence band do make a small contribution to the current passing through the semiconductor; however, theirs is a minority role.

In p-type semiconductors, the absence of a fourth valence electron in the Group III atom leaves one of the valence electrons in some of the silicon atoms unpaired (Fig 6.32c).

To offset this deficiency, an electron from a nearby covalent bond may jump across and fill the gap.

Since, overall, this action produces no change in the lattice bonding, it requires just a small amount of energy, E_A, typically -0.05eV.

This amount of energy is readily obtained from the thermal vibrations of the lattice.

However, the volunteer electron leaves a positive ion at the position it vacates; this is equivalent to the creation of a hole in the valance band.

The hole can be filled by the transfer of another electron from a different covalent bond, but this in tum just creates a new positive hole somewhere else.

Thus, the net result of introducing Group III impurity atoms into the silicon lattice is a random motion of positive holes from one lattice site to another.

This random motion of the positive holes in p-type semiconductors is analogous to that of the conduction electrons in n-type semiconductors.

In terms of energy levels, the valence electrons flitting between the positive holes in the semiconductor material occupy an acceptor energy level that is just above the valence band, i.e., $E_A << E_g$,

This increases the number of holes in the valence band and so enhances the material's conductivity.

However, the number of electrons in the conduction band is unaffected (Fig 6.33).

It follows that the principle (majority) charge carriers in p-type semiconductors are the holes in the valence band; the electrons in the conduction band are the minority charge carriers.

6.4.8 The p-n Junction

The extent to which an extrinsic semiconductor's conductivity is enhanced closely depends on the concentration of the impurity.

Techniques exist for manufacturing single semiconductor crystals with precisely determined impurity concentrations.

It is even possible to produce single crystals one end of which is an n-type material and the other a p-type; such crystals are called p-n Junction Diodes.

The moment a junction diode is formed (Fig 6.34a), majority charge carriers start diffusing across the line between the two materials, i.e., electrons diffuse from the n-type material into the p-type material and holes from the p-type into the n-type.

This exchange of majority charge carriers is called the diffusion current.

It effectively transfers positive charge from the p-type side of the seam to the n-type side, thereby establishing a contact potential difference, V_0, of several tenths of a volt across the boundary between them (Fig 6.34b).

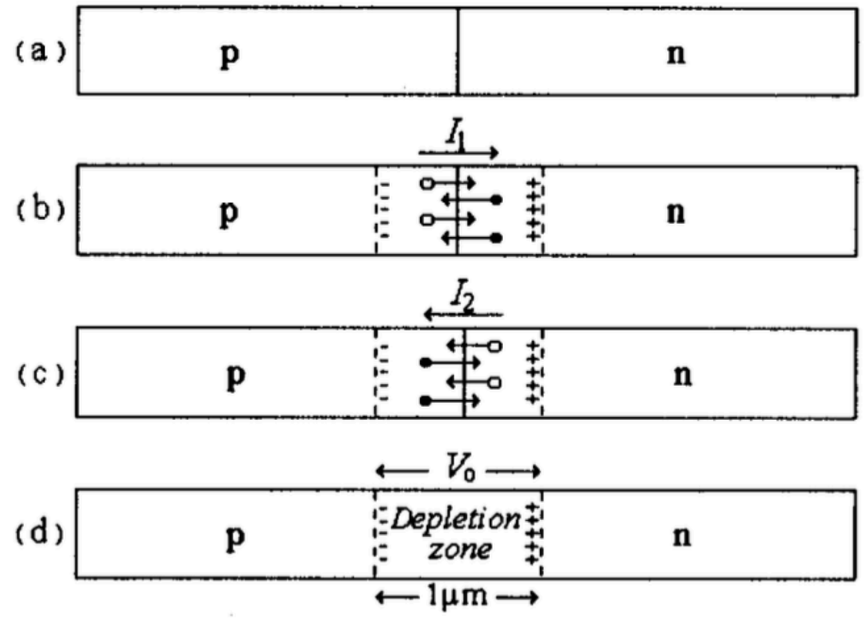


Fig 6.34 a. A p-n junction diode. b. The diffusion current, I_1, of the majority charge carriers. c. The drift current, I_2, of the minority charge carriers. d. The contact potential, V_0, set up across the depletion zone.

Under the influence of the contact potential, minority charge carriers now start to move in the opposite direction across the narrow seam joining the two materials; holes pass from the n-type material in to the p-type and electrons from the p-type into the n-type (Fig 6.34c).

This exchange of minority charge carriers, called the drift current, effectively transfers positive charge from the n-type side of the seam to the p-type side.

At equilibrium, the rate, I_1, at which positive charge is transferred from the p-type material to the n-type by the diffusion of the majority charge carriers is balanced by the rate, I_2 at which it is returned by the drift of the minority charge carriers from the n-type material to the p-type, i.e.,

$$I_1 = I_2 \tag{6.58}$$

The exchange of charge carriers takes place in a narrow strip of material, typically of width ~ lμm, that overlaps the boundary between the two materials on both sides.

This strip, called the depletion zone, is effectively emptied of charge carriers.

The only charged entities remaining in the depletion zone are the positive ions of the Group V atoms in the n-type material that have lost their fifth valence electron and the negative ions of the Group III atoms in the p-type material that have accepted these electrons; all the holes have been filled.

These ions are locked in the lattice and so are immobile; it is they that maintain the contact potential at equilibrium.

In terms of energy levels, the formation of a junction diode involves the merging of the conduction and valence bands of the two materials across the boundary between them.

This enables the relatively numerous electrons in the n-type material's conduction band and the relatively numerous positive holes in the p-type material's valence bond to migrate across the junction and neutralize one another (Fig 6.35a).

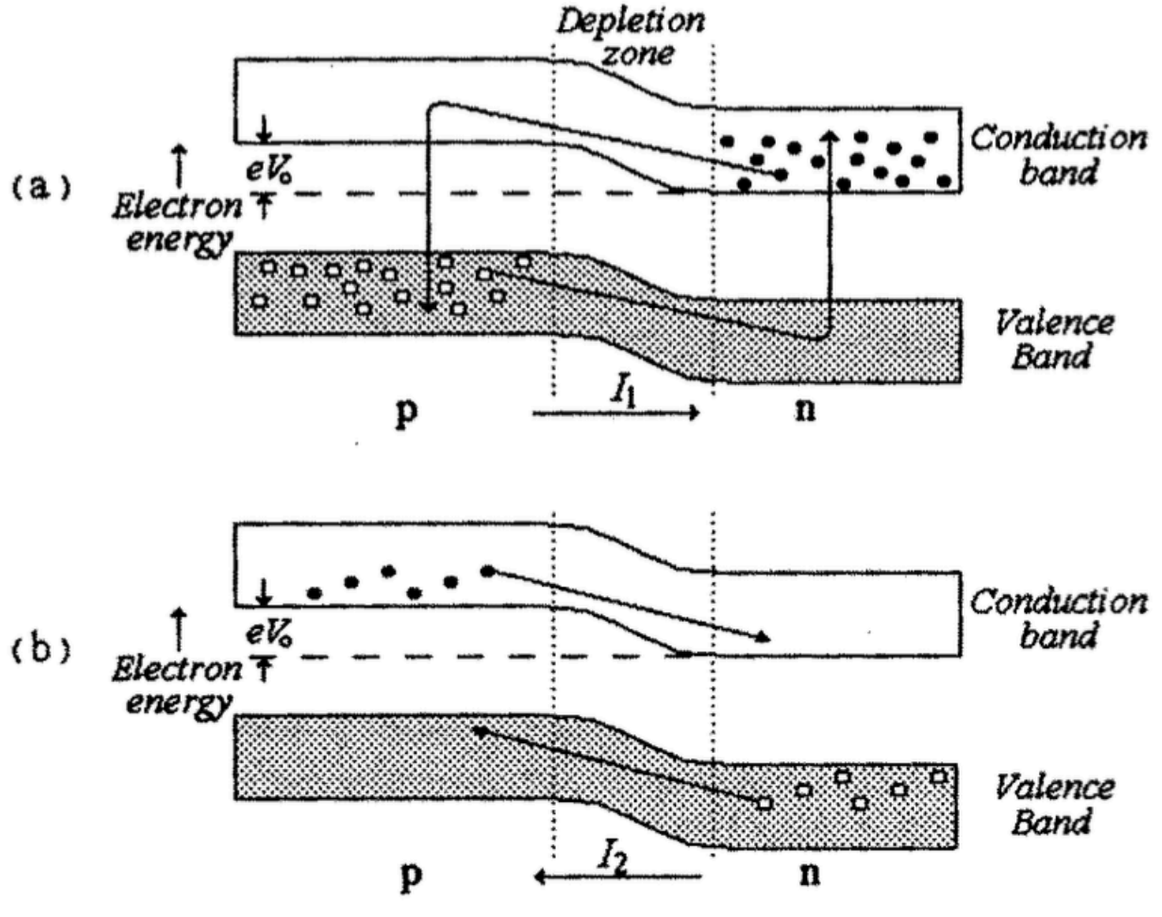


Fig 6.35 The electronic energy level diagram of an unbiased p-n junction. (a) The diffusion current of the majority charge carriers. Electrons in the conduction band migrate fromthe n-type material into the p-type where they fall into the valence band and neutralize positive holes. Simultaneously, holes in the valence band migrate in the opposite direction and 'fall' into the conduction band where they are neutralized by electrons.(b) The drift current of the minority charge carriers.

The potential difference thus set up across the junction lowers the electronic energy levels on the n-type side relative to the p-type side by an amount eV_0.

This energy difference causes electrons in the conduction band to drift across the junction from the p-type material and holes in the valence band to drift in the opposite direction (Fig 6.35b).

The potential difference across the boundary in a p-n junction can be modulated by means of an external potential source, *V*.

By connecting the positive terminal of the external source to the n-type end of the junction and the negative terminal to the p-type end, the potential is raised to (V_O+V).

This arrangement, called a *reverse bias connection*, has the effect of significantly reducing the diffusion current, I_1, the thermal vibrations of the lattice being too weak at normal temperatures to provide the majority charge carriers with the additional energy needed to overcome the increased potential.

On the other hand, the additional potential facilitates the drift current, I_2, and widens the depletion zone (Fig 6.36).

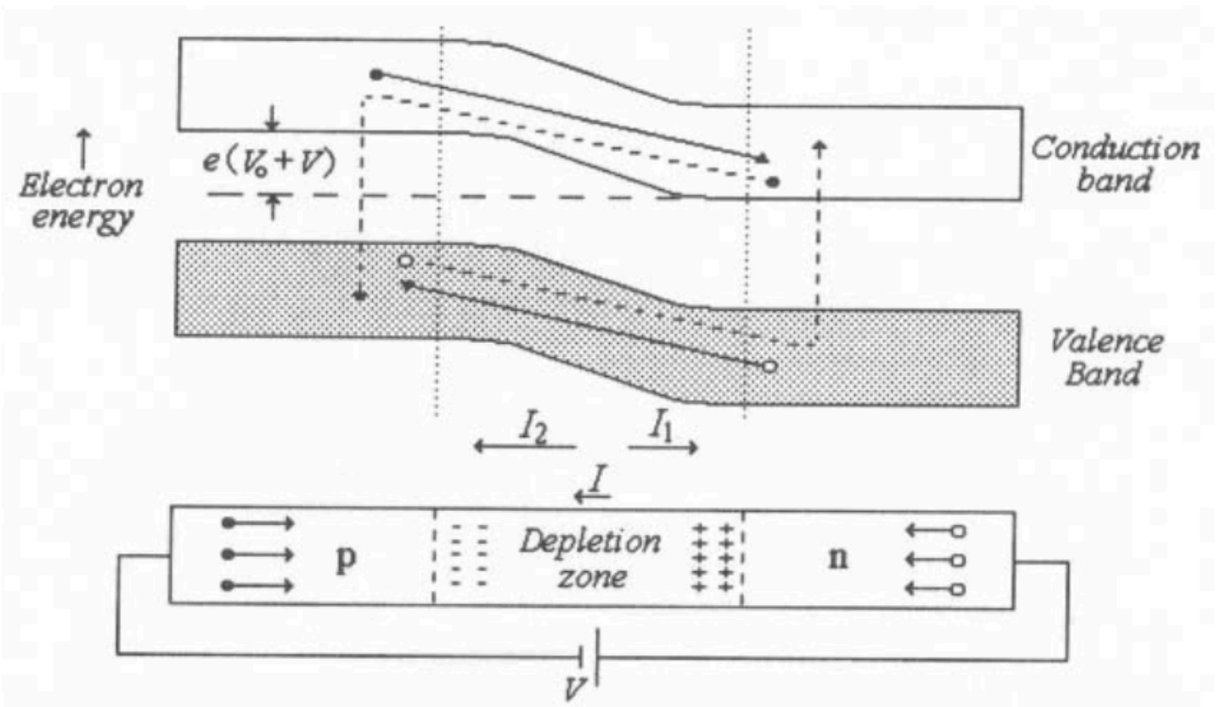


Fig 6.36 A reverse biased p-n junction. The diffusion current,I_1, is greatly reduced and is swamped by the drift current, I_2 As a result, a feeble current, $I = I_2 - I_1$, of minority charge carriers flows from the n-type to the p-type material.

The net current, $I = I_2 - I_1$, that flows through a reverse biased diode results from the drift of the minority charge carriers and is very weak, ~1μA.

The diode effectively exhibits a high resistance (>$10^5\Omega$) in the reverse bias connection that impedes the passage of any significant current.

When the positive terminal of the external source is connected to the p-type end of the junction and the negative terminal to the n-type end, the potential across the junction diode is reduced to $(V_0 - V)$.

This arrangement, called a *forward bias connection*, has the effect of significantly enhancing the diffusion current, I_1, of the majority charge carriers.

The reduced potential has little effect on the drift current, I_1 of the minority charge carriers though it does significantly narrow the depletion zone (Fig 6.37).

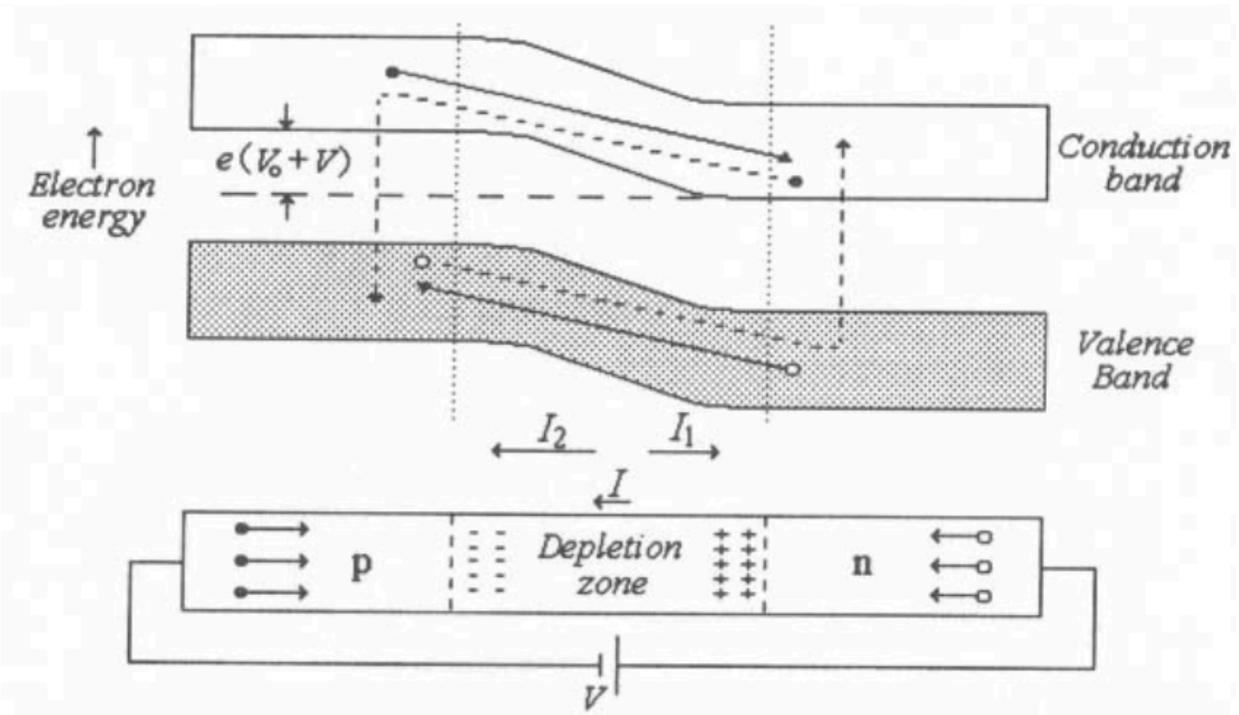


Fig 6.37 One obtains greatly enhanced diffusion current, I_1 of the majority charge carriers. The result is a net current, $I = I_2 - I_1$, whose direction is from the p-type to then-type material.

The net current, $I = I_2 - I_1$, flowing through a forward biased diode results from the injection of majority charge carriers into and across the junction and is quite substantial, ~l mA.

Thus, in the forward bias connection the diode effectively exhibits a relatively low resistance (~100 -10000Ω).

In general, the magnitude of the diffusion current, I_1, depends on the number of majority charge carriers that, at a given absolute temperature, T, possess sufficient energy to overcome the potential barrier, $(V_0 \pm V)$, between the two ends of the p-n junction.

Assuming a Maxwell-Boltzmann distribution of the energies of the majority charge carriers, this gives:

$$I_1 = Ae^{-(V_0 \pm V)/kT} \tag{6.59}$$

where A is a constant of the system.

At equilibrium, when V = 0, the diffusion current equals the drift current, i.e., $I_1 = I_2$. Thus,

$$I_2 = Ae^{-V_0/kT} \tag{6.60}$$

Accordingly, the current, I, through a junction diode is given by the function:

$$\begin{aligned} I &= I_1 - I_2 \\ &= Ae^{-(V_0 \pm V)/kT} - Ae^{-V_0/kT} \\ &= I_2\left(e^{\pm V/kT} - 1\right) \end{aligned} \tag{6.61}$$

which, to a good approximation, matches the experimental diode characteristic shown in Fig 6.38.

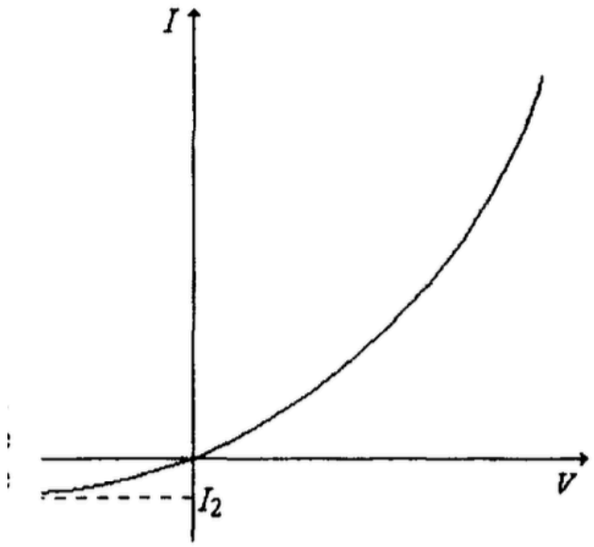


Fig 6.38 The characteristic of a junction diode. The diode exhibits excellent conductivity in one direction but poor conductivity in the opposite direction.

6.4.9 Semiconductor Devices

The three most important functional properties of p-n junction diodes are:

(i) they are excellent conductors of electricity in one direction but not in the other;

(ii) their interactions with radiation are similar to those of ordinary atoms;

(iii) they react almost instantaneously to electrical changes.

Moreover, by the techniques of miniaturization employed in the production of integrated circuits. p-n junctions can be made microscopically small.

It is for these reasons that junction diodes are the principle constituents of most modern electronic devices.

The electronic and computer revolution of the late twentieth century could never have taken place without the development of p-n junction diodes and the semiconductor devices such as the transistors and photodiodes that are based upon it.

1. Photodiodes

Photons of sufficient energy that strike an unbiased p-n junction may be absorbed by electrons in the semiconductor's valence bond.

The electrons are excited into the conduction band leaving positive holes in the valence band.

The net result of the process is the annihilation of photons and the production of electron-hole pairs.

The contact potential across the depletion zone sweeps the electrons into the n-type material and the holes into the p-type material.

If the ends of the crystal are connected by a closed circuit, a current will flow, and the irradiated junction acts as a source of EMF (Fig 6.39).

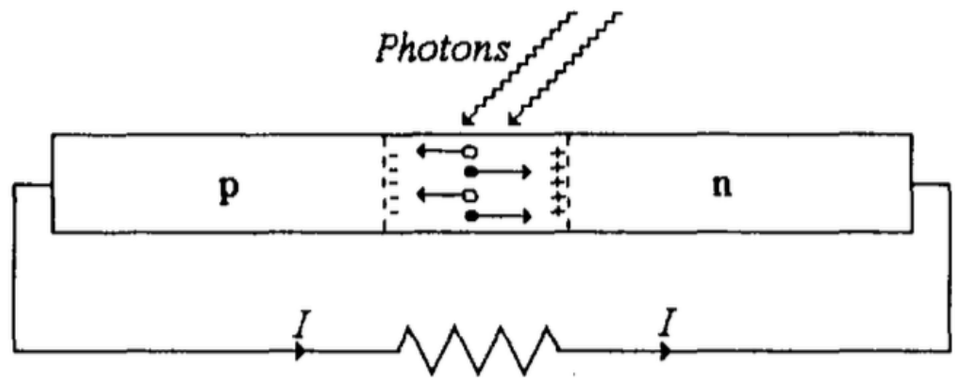


Fig 6.39 A p-n junction operating as a source of EMF - a solar cell.

The solar cells, such as those in the panels that provide the electrical energy for space probes, typically comprise a thin transparent layer of p-type material deposited on a crystal of n-type material.

The EMF of such solar cells is ~ 0.6V. However, they only produce weak currents; a $5cm^2$ silicon based solar cell produces just 0.1A in full sunlight.

Solar cells convert about 10% of the incident radiation into electrical energy.

Electron-hole pairs are also produced when an ionizing particle passes through the depletion zone of a reverse biased p-n junction.

Each pair results in a short pulse of current.

By counting these pulses, the intensity of the radiation can be measured.

This is the basis for the use of p-n junctions as radiation detectors.

Just as p-n junction diodes absorb radiation, they can also be made to emit radiation.

Although the energy released when an electron falls from the conduction band into the valance band and combines with a positive hole usually appears as heat, in forward biased junction diodes made out of heavily doped semiconductor materials such as gallium arsenide, some of this energy may be released as radiation of frequency $f = E_g/\hbar$.

Typically, the energy difference, E_g, between the conduction band and the valence band in semiconductors is in the range 0.5 to 3.0eV.

Photons of infra-red and visible light possess energies in this range and so the diodes emit radiation in the infra-red and visible regions of the spectrum.

These light emitting diodes have many uses and they are both reliable and cheap to produce.

They are the small red lights typically used to indicate that a radio receiver or a component in a computer such as a disk drive is in operation.

Light emitting diodes exhibit both spontaneous and stimulated emission.

The injection of electrons into the conduction band on the p-side of a forward biased p-n junction and of holes into the valence band on the n-side establishes a population inversion across its plane; large numbers of electrons enter the p-region of the diode and holes enter the n-region.

The greater the current flowing through the junction, the stronger the inversion.

At low currents, i.e., modest population inversions, spontaneous emission dominates. At high currents, i.e., strong population inversions, stimulated emission dominates.

Thus, a suitably proportioned and pumped light emitting diode *can* function as a laser (Fig 6.40)

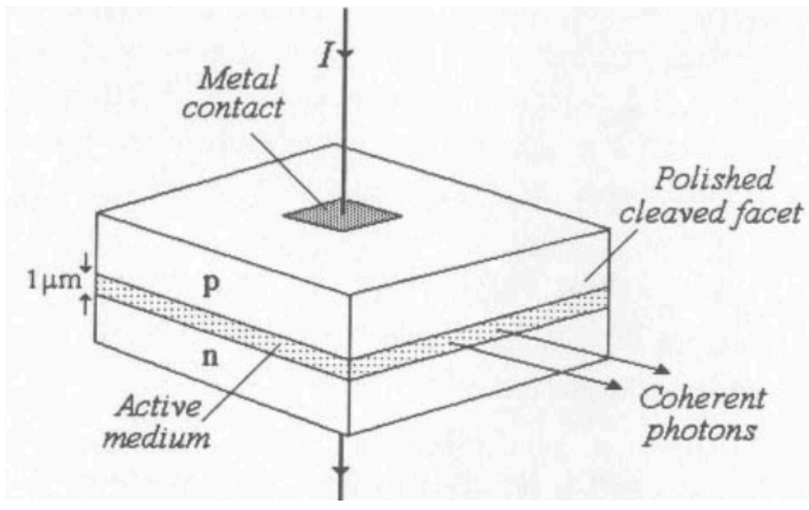


Fig 6.40 Typical layered structure of a gallium arsenide laser diode. The laser is pumped by the current, I. The active medium acts as a resonator.

In their simplest form, laser diodes are small rectangular pieces of semiconducting material with two cleaved facets at opposite ends that act as mirrors.

The active medium in which the lasing occurs is sandwiched between a layer of p-type material and a layer of n-type material.

The beam emitted at the output facet is typically just a few micrometers wide.

Beam wavelengths are generally between 780nm and 905nm and their intensities between 10mW and 25mW.

Laser diodes are both robust and highly efficient (25% - 30%).

Laser diodes operate at modest currents, ~20mA, and react almost instantaneously to modulations in the pumping current.

Thus, laser diodes can be repeatedly and rapidly switched on and off; it is this property that underlies their widespread use, in conjunction with optical fibers, by the telecommunications industry.

2. Junction transistors

The junction (bipolar) transistor was invented by Bardeen, Brattain and Shockley in 1948.

The essential properties of a junction transistor are its capabilities to amplify an electrical signal and to switch a current on or off almost instantaneously.

The junction transistor is a semiconductor crystal, composed of three layers of doped materials, either p-n-p or n-p-n.

This arrangement effectively comprises two p-n junctions connected back to back (Fig 6.41).

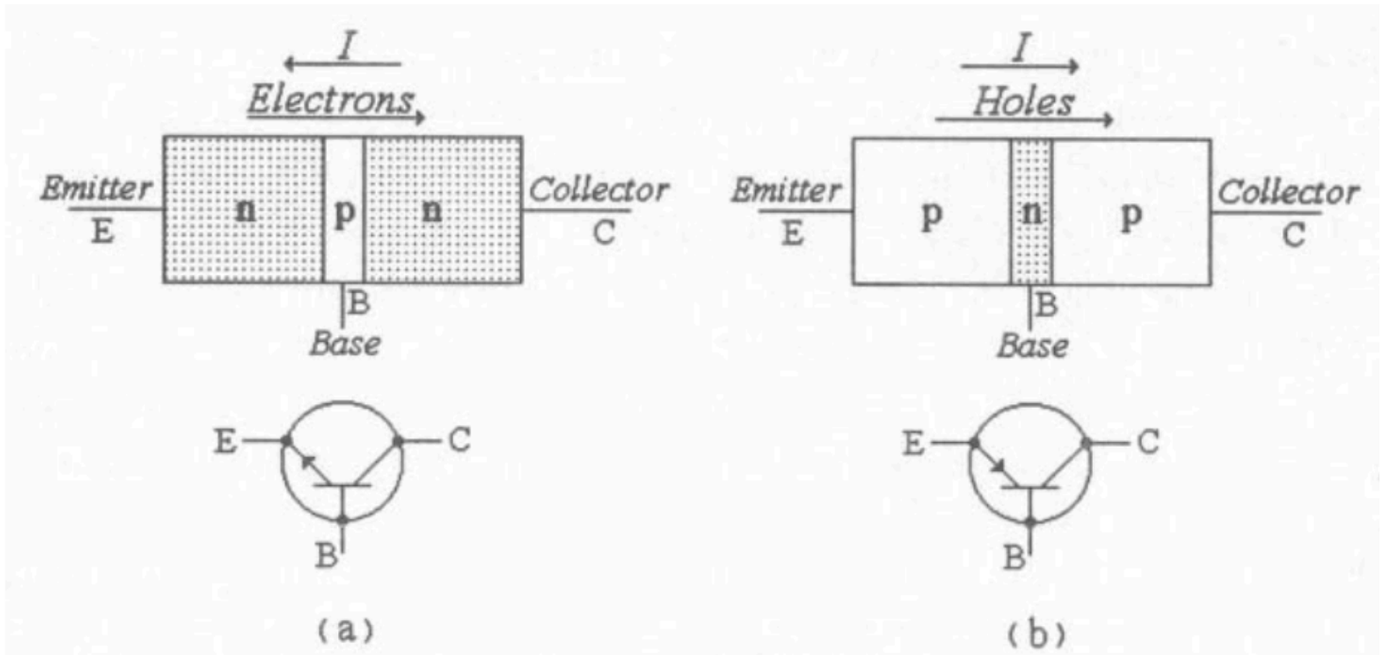


Fig 6.41 The junction transistor. Transistors are three terminal devices. (a) An n-p-n transistor. The majority charge carriers are the electrons that flow from the emitter to the collector in the opposite direction to that of the current. (b) A p-n-p transistor. The majority charge carriers are the holes that flow from the emitter to the collector in the direction of the current.

The three layers of material constituting the transistor are called the emitter, base and collector, respectively.

The majority charge carriers in p-n-p transistors are holes; the majority charge carriers in n-p-n transistors are electrons.

The base is very thin, ~1μm, and is usually made of more lightly doped material than the emitter and collector.

Junction transistors are most usually connected in a circuit with the p-n junction between the emitter and the base forward biased and that between the base and the collector reverse biased (Fig 6.42).

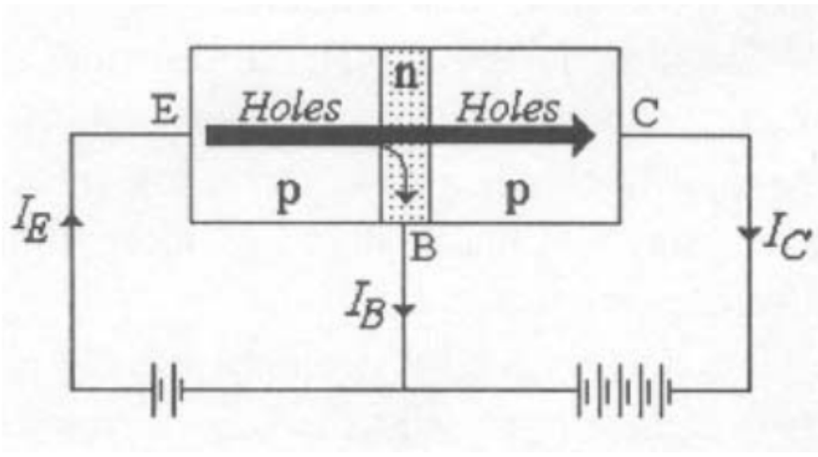


Fig 6.42 The flow of the majority charge carriers -the holes- and the currents through a p-n-p transistor.

As a result, a relatively large number of majority charge carriers flow across the junction from the emitter into the base.

This flow of majority charge carriers constitutes the *emitter current*, I_E.

Because of the thinness of the base and the sparsity of its impurity atoms, most of the holes entering the base from a p-type emitter (or the electrons entering from an n-type emitter) do not encounter electrons (or holes) with which they can combine.

Since both the emitter and the collector share the same type of majority charge carrier and the p-n junction between the base and the collector is reverse biased, i.e., it encourages the flow of the collector's majority charge carriers from the base into the collector, most of the emitter's majority charge carriers entering the base pass straight through it and are swept up into the collector.

This transfer occurs despite the fact that reverse biased p-n junctions usually exhibit a high resistance (often $>10^5\Omega$). The name transistor was derived from this 'transfer-resistor' effect.

The flow of majority charge carriers into the collector constitutes the *collector current*, I_C.

The small number of charge carriers entering the base from the emitter that do encounter an electron (or hole, as the case may be), or that leak out of the base and back into the emitter constitute the *base current*, I_B.

The relation between the three currents, I_E, I_C and I_B, is given by:

$$I_C = I_E - I_B \tag{6.62}$$

The ratio, $\beta = I_C / I_B$, between the collector current and the base current is a characteristic parameter of junction transistors.

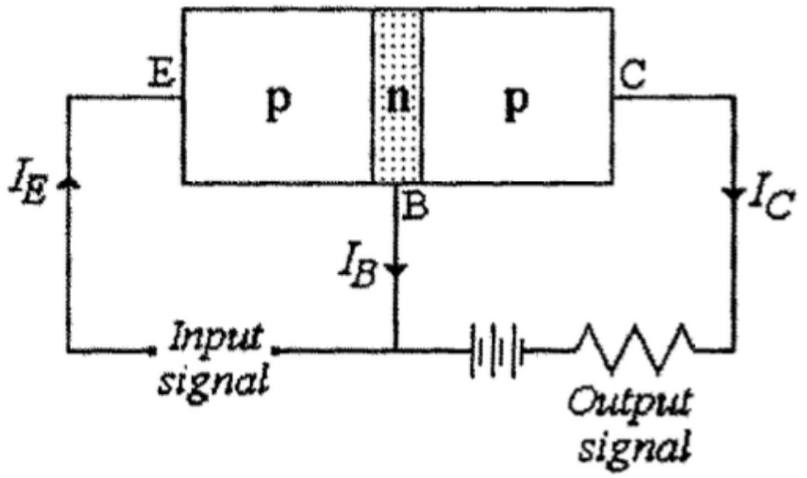
Its value depends on the geometry of the transistor and the composition of its component layers.

Typical values for β vary between 20 and 200.

Thus, even weak base currents can produce considerable collector currents, small fluctuations in the former appearing as large variations in the latter.

It is this effect that underlies the use of transistors in amplifying circuits (Fig 6.43).

Fig 6.43 A simple transistor amplifier circuit



In the simplest arrangement, the signal to be amplified is applied across the p-n junction between the emitter and the base and constitutes its forward biasing potential.

The collector current multiplies the resulting base current by the factor β. Any modulations in the signal are almost instantaneously reflected by variations in the collector current with very little distortion.

The power of the output signal, P_{Output}, depends not only on the collector current, I_C, but also on the voltage, V_{BC}, in the collector-base circuit.

$$P_{Output} = I_C V_{BC} \tag{6.63}$$

Since the p-n junction in this circuit is reverse biased, substantial collector-base voltages can be used.

For example, the collector-base voltage, V_{BC}, can typically be 200 times the emitter-base voltage, V_{BE}.

This has the effect of further multiplying the power of the output.

$$
\begin{aligned}
P_{Output} &= I_C \cdot 200 V_{BE} \\
&= \beta I_B \cdot 200 V_{BE} \\
&= 200 \beta P_{Input}
\end{aligned}
\tag{6.64}
$$

where $P_{Input} = I_B V_{BE}$. In this way a current ratio of $\beta \approx 100$ can produce a power amplification >20,000.

3. Field effect transistors

Whereas the magnitude of the current flowing through junction transistors is determined by the stock of majority charge carriers provided by the base current, in field effect transistors it is the effective cross-sectional area of the semiconductor channel through which the current flows that modulates its magnitude.

Like the junction transistor, the *field-effect transistor* - FET is also a three terminal device.

In the FET, the terminals are denoted the *source*, *drain* and *gate* and they correspond, respectively, to the emitter, collector and base of the junction diode.

Of the many different types of FET's we will describe the MOSFET - *Metal Oxide Semiconductor Field Effect Transistor* (Fig 6.44).

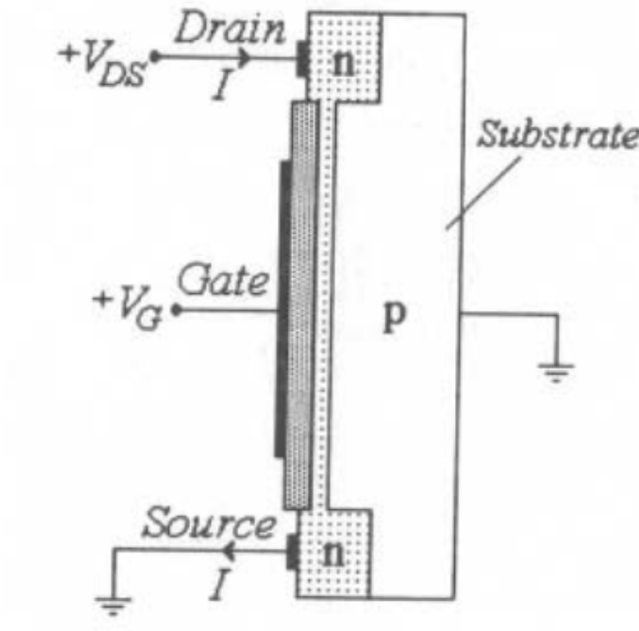


Fig 6.44 An n-channel MOSFET (Metal Oxide Semiconductor Field Effect Transistor). The p-n junction is reverse biased by the gate-substrate potential, V_G. This results in the formation of a depletion zone extending the length of the narrow channel of n-type material that links the drain and the source. Whether or not a current, I, flows between the drain and the source, depends on the width of the depletion zone; this in tum depends on the magnitude of the gate-substrate potential.

Two nuggets of heavily doped n-type semiconductor material, each of which has a metal terminal attached to it, are embedded into the ends of a substrate of lightly doped p-type material.

The metal terminals are the drain and the source of the transistor.

A shallow layer of n-type material, called the n-channel, is then deposited onto the surface of the substrate between the nuggets so as to link them into an unbroken section of n-type semiconductor.

This creates a p-n junction along the whole length of the channel.

A thin coating of an oxide insulator is now laid over the n-channel.

This, in turn, is plated with a metallic film.

A terminal is attached to this metal film and constitutes the transistor's gate terminal.

Note that the gate and the semiconductor materials are insulated from one another; no current can flow between them.

In operation, two potentials are applied to the MOSFET.

The first, VDS, is applied between the drain and the source; this causes a current, I, to flow down the n-channel.

The majority charge carriers in the n-channel are electrons and these flow from the source to the drain (against the conventional direction of the electric current).

The second potential, V_0, is applied between gate and the substrate of p-type material.

Because of the layer of insulating oxide between them, no current flows between the gate and the substrate.

However, an electric field is established between them and this has the effect of reverse biasing the p-n junction along the channel and of widening the depletion zone along its length.

Since the depletion zone, as its name implies, is empty of charge carriers, the gate-substrate potential, V_0, effectively determines the cross-sectional area of the n-channel through which current can flow, i.e., its conductivity.

If the gate-substrate potential is sufficiently strong, typically ~5V, then at some point along its length the depletion zone will be wider than the channel.

Under these circumstances no electric current can flow along its length; the transistor is said to be “OFF".

If a current does flow, i.e., the gate base potential is too small for the depletion zone to expand over the width of the channel, typically ~0.5V, the transistor is said to be "ON".

Field-effect transistors are particularly suitable for use in electronic computers.

The ‘0’ of the binary arithmetic on which their logic is based can be made to correspond to the transistor being OFF (gate closed) and the '1' to the transistor being ON (gate open).

FET's can switch between these two modes almost instantaneously.

Moreover, since the operation of the gate draws no current from the circuit, it does not corrupt or affect the information in the signals being exchanged between the computer's components.

Moreover, FET's are relatively simple to make and are readily incorporated into the Very Large Scale Integrated (VLSI) circuits that are at the heart of modern computers.

Their power usage is minimal and they have proved to be both robust and reliable under a range of operating conditions.

From the chips in personal computers to the central processing units of super computers, they have become the indispensable core of the modem world.

Part 6.5 Invariance, Symmetry and Conservation Laws

Symmetry is a familiar concept.

Examples of it abound in Nature: snowflakes, crystals and vertebrate animals all exhibit certain symmetries.

In general, the invariant features of a system, i.e., those that remain unchanged after some operation has been performed upon them, constitute its symmetry.

The simplest type of symmetry is the bilateral symmetry we observe when reflection about a line leaves the form of the object unaltered (Fig 6.45a). Another type is the rotational symmetry exhibited by the pentagram; rotations around its centre through angles that are integral multiples of 360°/5 leave the figure unaltered (6.45b).

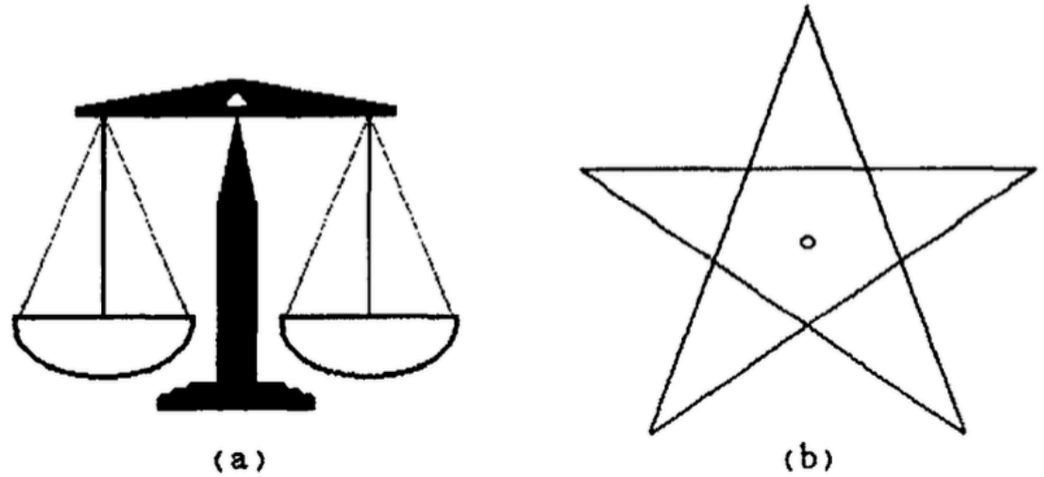


Fig 6.45 Examples of symmetry. (a) Bilateral symmetry: reflection about a line drawn down the central column leaves the figure unaltered. (b) Rotational symmetry: the pentagram is carried over into itself by rotations around its center through angles that are an integer multiple of 360°/5. The pentagram also exhibits bilateral symmetry.

Considerations of symmetry are often of great use in physics.

The motion of a simple pendulum (or of any simple harmonic motion) is symmetrical.

Quantities such as the system's velocity and acceleration have the same magnitude at symmetric positions (Fig 6.46).

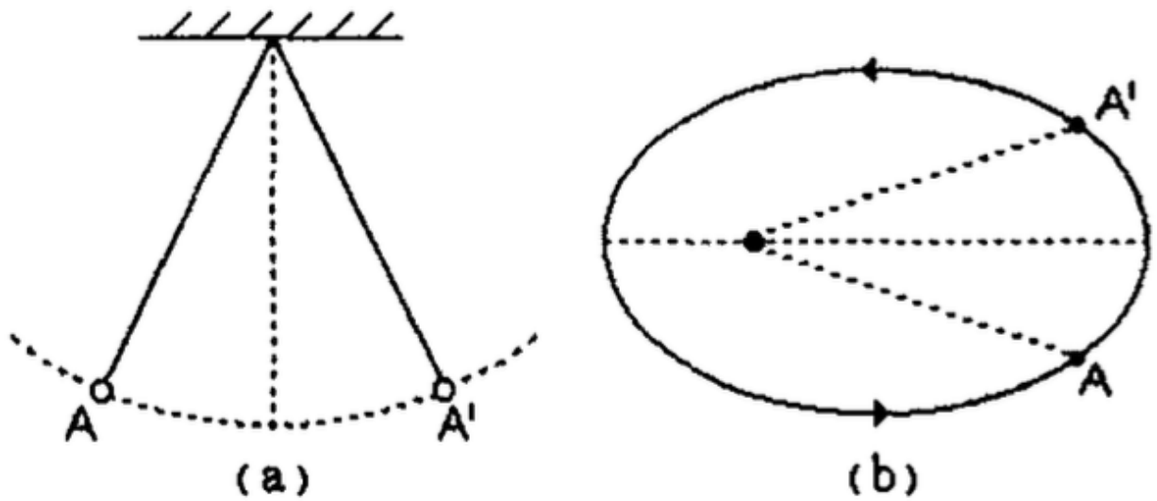


Fig 6.46 Symmetrical motions. (a) A pendulum. (b) A planetary orbit. The points A and A' are symmetric.

Likewise, the system's potential energy is symmetric relative to the central equilibrium point.

In a quantum system, the probability densities of two symmetrical points, must be equal: $|\psi_A|^2 = |\psi_{A'}|^2$.

However, what concerns us here is not the symmetry of physical systems, objects or geometrical shapes but that of the fundamental laws which direct the workings of the physical world.

6.5.1 The Symmetry of the Laws of Physics

In modern times, the recognition of the importance of symmetry in understanding the physical world began largely with Einstein's theories of relativity.

Underlying these theories was Einstein's conviction that the laws of physics must take the same form in all reference frames.

The laws must be covariant and suitable transformations must be found and formulated to ensure this.

This axiom incorporated the concept of observer equivalence, that was implicit in the belief in a physical reality that is independent of any percipient entity.

In symmetry terms, the principle of covariance requires that the operation of moving from one reference frame to another must not alter the laws of physics.

Like the appearance of the pentagram, the form of the laws must remain unaltered when certain operations are performed upon them.

Those operations which do not change the form of the laws are their symmetry operations.

They correspond to the transformations of the conventional approach to relativity.

Important symmetries other than those imputed by Einstein, had already been incorporated into scientific thinking when he published his theories of relativity.

The symmetry of Nature under translations in time, however distant that time may be, was first expressed by Hutton in the Principle of Uniformitarianism, which asserts that all geologic phenomena may be explained as the result of existing forces having operated uniformly from the origin of the earth to the present time.

It is the belief in this symmetry implicit in modern theories of evolution and cosmology that distinguishes them from creation myths.

The notion of identical elementary particles, whether they are atoms of a certain element or the electrons, protons and neutrons that comprise these atoms, expresses another kind of symmetry.

It is a symmetry under exchange.

Replacing an oxygen atom in a molecule by another oxygen atom makes no difference to the chemical properties of the molecule all oxygen atoms are chemically identical.

Time reversal, the operation usually designated by the symbol T, is a perplexing operation.

Whereas it can be easily shown that the laws of classical mechanics and electromagnetism are invariant under time reversal, i.e., if t is replaced by -t wherever it appears in the equations, our intuition and daily experience point to strong evidence that time flows in just one direction, from the past to the future.

In physical terms, the effect of time reversal is to exchange the final and initial states.

Like running a motion picture backwards; starting with the fried egg and finishing with a freshly laid egg.

As far as we can tell, this macroscopic irreversibility is a consequence of the very large number of particles involved.

It underlies the thermodynamic concept that the entropy of the physical universe is always rising.

But if we look at individual atoms we find that the laws that govern their behavior are time reversible.

Not that the laws of Nature are all invariably symmetrical - for example, their asymmetry under changes of scale was first noted by Galileo, when he recognized that:

Similar solid objects do not exhibit a strength that is proportional to their size....A large beam or column falling from a given height will go to pieces, when, under the same circumstances, a small scantling or small marble cylinder will not break....Clearly then, if one wishes to maintain in a great giant the same proportion of limb as that found in an ordinary man, he must either find a harder and stronger material for making the bones or he must admit a diminution of strength....Whereas, if the size of a body be diminished, the strength of that body is not diminished in the same proportion; indeed the smaller the body the greater its relative strength.

By providing a simple rational explanation for so many familiar natural phenomena, this simple inference did perhaps even more than Galileo's astronomical observations to undermine the traditional and hallowed beliefs of his time.

6.5.2 Group Theory

The notion that the concept of symmetry can be rigorously applied to abstracts such as the laws of Nature is quite new.

It is based on the concept of a *group* introduced into mathematics by Galois.

A group is any set for which a rule, symbolized by *, can be defined such that the product, a*b, of combining any two elements of the set is itself a member of the set.

For instance, the set of integers

$$\ldots, -3, -2, -1, 0, +1, +2, +3, \ldots$$

forms a group with the combination rule 'addition'.

As required, the product of combining any two integers by addition gives another integer:

$$(-3)+(+2)=(-1) \tag{6.65}$$

In this case, the symbol * is replaced by+. Such a group is called an *additive group*.

There are three other conditions that must also be satisfied by a group:

1. **The product of combining a particular element of the set, called the *identity*, with any other element of the set must leave the latter unchanged: i.e., $a^*I = a$, where I symbolises the identity. In the additive group, the zero fulfils this role:**

$$n + 0 = n \tag{6.66}$$

2. **Every element, a, has an inverse element, a^{-1}, such that their combination equals the identity: i.e., a^*a^{-1} = *the identity*. In the additive group, for each integer, n, there is a negative integer, $-n$, such that their sum equals zero, and *vice versa*.**

$$n + (-n) = 0 \tag{6.67}$$

3. **The combination rule must be associative, i.e., a*(b*c) = (a*b)*c. This is satisfied by the additive group:**

$$(+2)+[(-3)+(-1)]=-2=[(+2)+(-3)]+(-1) \tag{6.68}$$

Although the rule must be associative, it is not necessarily commutative.

That is a*b need not equal b*a. A group, all of whose elements commute, is called an Abelian group; the additive integer group is Abelian:

$$[(-3)+(-1)] = -4 = [(-1)+(-3)] \tag{6.69}$$

Non-commutative groups are termed *non-Abelian*.

Many different entities and operations can constitute a group.

For example, vectors form an additive group similar to the additive integer group, though the operation performed when adding vectors is different (Fig 6.47).

But other than that, the rules for dealing with vectors are exactly the same as those for dealing with integers.

The zero vector is a vector of zero magnitude; the inverse is a vector in the opposite direction.

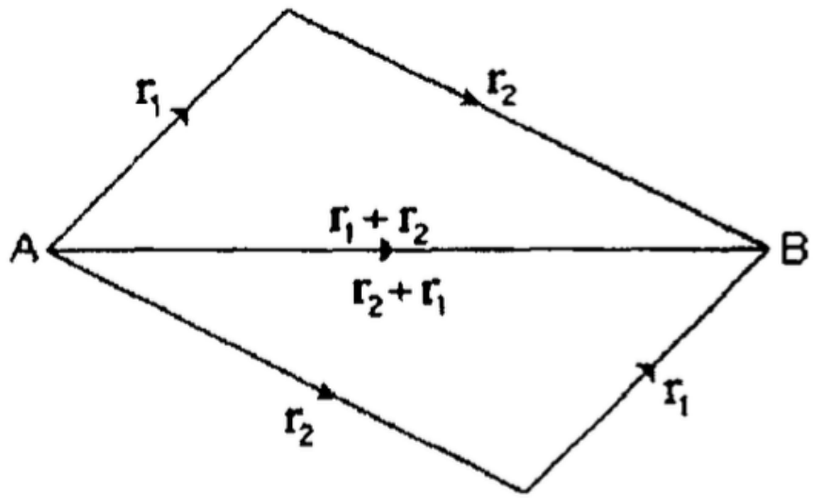


Fig 6.47 The displacement from A to B is itself a vector which can be defined as the sums $\mathbf{r}_1 + \mathbf{r}_2$ or $\mathbf{r}_2 + \mathbf{r}_1$. From the diagram it is clear that $\mathbf{r}_1 + \mathbf{r}_2 = \mathbf{r}_2 + \mathbf{r}_1$.

Another example of an additive group is the rotations of a line in a given plane (Fig 6.48). The zero of this system represents no rotation. The inverse is a rotation in the opposite direction.

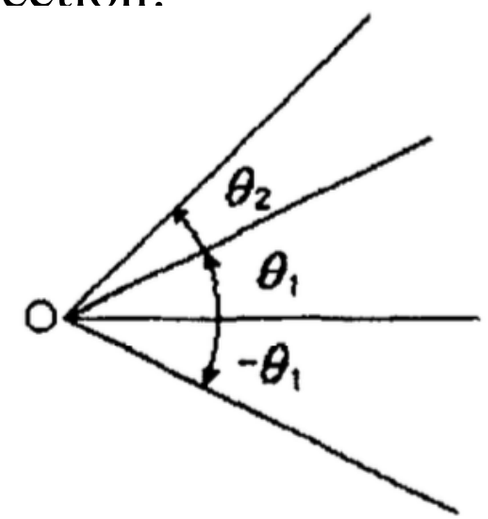


Fig 6.48 Rotations in a plane about the point O. Rotations are positive if they are anti-clockwise and negative if clockwise. The rotations are commutative because evidently $\theta_1 + \theta_2 = \theta_2 + \theta_1$.

The integer, vector and rotation groups we have so far considered are all examples of *infinite groups*; they all contain an infinite number of elements.

A simple example of a *finite group* is given by the symmetric rotations of the three pronged shape-tripod shown in Fig 6.49.

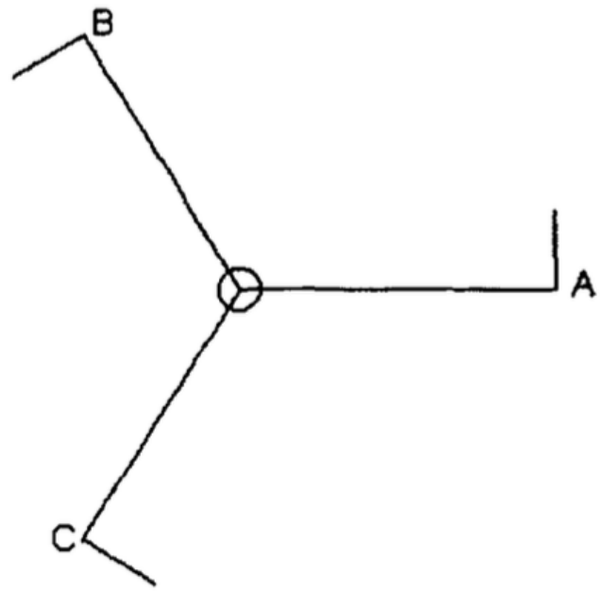


Fig 6.49 Rotational symmetry. There are just three rotations around the tripod's centre which leave its appearance unchanged: an anticlockwise rotation through 120°; an anticlockwise rotation through 240°; an anticlockwise rotation through 360°, which amounts to no rotation at all or to C leaving things as they were.

There are just three rotations around its center which leave its appearance unchanged: an anticlockwise rotation through 120°, which we designate $\mathbf{Rot}_{+120°}$; an anticlockwise rotation through 240°, which we designate $\mathbf{Rot}_{+240°}$; an anticlockwise rotation through 360°, which we designate the identity, **I** ,as it leaves things as they were.

We can disregard clockwise rotations of the tripod as they are indistinguishable from the anti-clockwise rotations:

$$\mathbf{Rot}_{+120^\circ} = \mathbf{Rot}_{-240^\circ} \text{ and } \mathbf{Rot}_{+240^\circ} = \mathbf{Rot}_{-120^\circ} \tag{6.70}$$

The combination rule for this symmetry group is 'after applying the first operation, apply the second one'.

Applying this rule we see that two successive rotations through 120° have the same effect as one rotation through 240°. We can express this symbolically as:

$$\mathbf{Rot}_{+120^\circ} * \mathbf{Rot}_{+120^\circ} = \mathbf{Rot}_{+240^\circ} \tag{6.71}$$

Similarly, two 240° are equivalent to one 120° rotation:

$$\mathbf{Rot}_{+240^\circ} * \mathbf{Rot}_{+240^\circ} = \mathbf{Rot}_{+120^\circ} \tag{6.72}$$

The group of rotational symmetry operations that can be performed on the tripod can be summarized in a matrix termed a *multiplication table* in group theory.

*	I	$\mathbf{Rot}_{+120^\circ}$	$\mathbf{Rot}_{+240^\circ}$
I	I	$\mathbf{Rot}_{+120^\circ}$	$\mathbf{Rot}_{+240^\circ}$
$\mathbf{Rot}_{+120^\circ}$	$\mathbf{Rot}_{+120^\circ}$	$\mathbf{Rot}_{+240^\circ}$	I
$\mathbf{Rot}_{+240^\circ}$	$\mathbf{Rot}_{+240^\circ}$	I	$\mathbf{Rot}_{+120^\circ}$

An examination of the table shows that this group is commutative (Abelian). For example:

$$\mathbf{Rot}_{+240^\circ} * \mathbf{Rot}_{+240^\circ} = I = \mathbf{Rot}_{+120^\circ} * \mathbf{Rot}_{+240^\circ} \tag{6.73}$$

Thus we see, that symmetry operations can be described in terms of an algebraic system. This is the key to the application of symmetry to the study of physical laws.

6.5.3 Noether's Theorem

A systematic study of the importance and significance of symmetry in the formulation of physical laws was only undertaken following the appearance of Einstein's theory of general relativity in 1915.

The incentive for this was a vexing problem that became immediately apparent with the publication of the theory, namely, that energy is not conserved locally in general relativity.

In classical field theories and in the theory of special relativity, it can be shown that the net outflow of energy across the boundaries of any arbitrary volume element is equal to the rate at which the energy inside the element decreases.

Physically, one has a localized conserved energy density. However, in general relativity, the localization of energy has no meaning.

A quantity can be defined that is analogous to the energy-momentum concept of special relativity but it is not invariant under general co-ordinate transformations different results are obtained in different reference frames.

In physical terms, there is a continuous transfer of energy to and from the all-pervading gravitational field which makes it meaningless to speak of the localization of the energy of the gravitational field.

Following a series of lectures on his new theory of general relativity given by Einstein at the University of Gottingen in 1915, this problem was taken up by Emmy Noether, who had just been invited to join the world-famous team of mathematicians led by David Hilbert.

In the landmark paper she published in 1918, Noether solved the problem, and in doing so, discovered general theorems concerning symmetry and the conservation of physical quantities that have profoundly affected modem physics.

In the first part of her paper, Noether revealed the general relationship that exists in field theories between symmetry and conservation:

To every symmetry there is a corresponding conservation law, and vice versa.

She then went on to show, in the second part of the paper, that the symmetry underlying the co-ordinate transformation of general relativity is different from that which underlies co-ordinate transformations in classical field theories and the theory of special relativity.

Hence, a quantity that is conserved in classical field theories will not necessarily be conserved in the same way in general relativity.

Prior to Noether's theorem, the conservation of energy was something of a mystery.

There were even those like Poincare who had contended that the conservation of energy is merely a definition of energy, since, whenever it appears to break down.

we simply invent a new form of energy to restore the balance.

Noether showed that energy conservation followed directly from symmetry under time translations; from the fact that the laws of Nature are the same on every day of the week.

Consider the law of gravity. Suppose the world was such that gravity was weaker on Mondays than on Tuesdays.

We could then raise objects on Mondays at a low expenditure of energy and make an energy profit by letting them fall back again on Tuesdays.

The process could be repeated week after week, thus increasing the store of energy in the universe as time went by.

The conservation of linear momentum follows from translational symmetry.

Consider a system of interacting particles and suppose that its potential energy, U, is independent of its position.

In this case, the derivatives of the potential energy with respect to the co-ordinates x, y and z would be zero, which implies that no external force is acting on the system: e.g. $F_x = - \partial U/\partial x = 0$.

Hence, the momentum of the system is unchanged: $\partial p_x/\partial t = Fx = 0$.

The conservation of angular momentum arises from the symmetry of the laws of physics under rotation of the co-ordinate system.

For a free object to suddenly start spinning of its own accord would certainly violate the law of angular momentum conservation.

It would also violate the symmetry of space for the spin axis would define a particular direction in space, distinguishable from all the others.

The symmetry of the classical laws of physics under translations in space and rotations of the co-ordinate axes is ensured by expressing them in terms of vectors.

If the vector equation $\mathbf{F = r}$ is true in one coordinate system, it is true in any coordinate system.

The great usefulness of vector relationships is that they carry their transformations within them; they are inherently symmetrical.

They incorporate the notion that empty space is the same in all directions.

6.5.4 The Conservation Laws of Particle Physics

The study of elementary particles has revealed new conservation laws and new symmetries.

Certain of these laws are obeyed under all circumstances; others are not.

Moreover, whereas some have arisen from theoretical considerations, others are just empirical relationships.

Notwithstanding, any reaction not expressly forbidden by the conservation laws will generally occur, if perhaps at a slow rate.

This expectation is based on quantum mechanics.

Unless the barrier between the initial and final states is infinitely high, there is always a non-zero probability that a system will make the transition between them.

In addition to the familiar conserved quantities listed above-energy, linear and angular momentum-those absolutely conserved in particle physics are (i) electric charge; (ii) lepton number; (iii) baryon number. Noether's theorem posits that there should be symmetries that with each of these conserved quantities too.

However, a corresponding symmetry has only been identified in the case of charge conservation.

(i) <u>The Conservation of Electric Charge</u>

The conservation of electric charge is associated with the symmetry called *gauge symmetry*.

A gauge transformation is one that has no observable consequence, like changing the arbitrary zero of the electric potential.

This has no effect on the electric field or on observable potential differences.

Like the height between two levels, the difference in potential between two points is independent of the base chosen for the zero.

Nor is the form of Maxwell's equations affected by the arbitrary choice of where the electrical potential is to equal zero.

It can be shown that charge conservation is a consequence of the gauge symmetry of Maxwell's equations.

Gauge symmetry is central to quantum mechanics and to the quantum field theories that underlie particle physics.

In quantum mechanics its meaning is quite straightforward.

The phase of a wave function, like the wave function itself, is unobservable.

It is only the probability density that is observable. However, if a wave function undergoes a phase transformation such that $\psi \rightarrow \psi e^{i\delta}$, the probability density is unaffected:

$$|\psi|^2 = |\psi e^{i\delta}|^2 \tag{6.74}$$

Thus the observables are conserved under gauge transformations.

(ii) The Conservation of Lepton Number

In essence, the conservation of lepton number means that whenever a lepton of a certain generation is created or destroyed in a reaction, a corresponding antilepton from the same generation must be created or destroyed.

The arithmetic is done by assigning elementary particles a lepton quantum number: +1 for the leptons, -1 for antileptons and zero for all other particles.

Consider the electron capture mode of decay.

The process involves only first generation leptons: electrons and neutrinos:

	p	+	e^-	$\rightarrow$	n	+	ν
Lepton Number	0		+1		0		+1

(6.75)

The particle emitted with the neutron must be a neutrino and not an antineutrino for lepton number to be conserved.

The conservation of lepton number has no clear theoretical basis nor has any simple symmetry yet been associated with it.

(iii) The Conservation of Baryon Number

Assigning elementary particles a baryon quantum number, +1 for the baryons, -1 for antibaryons and zero for all other particles enables us to formulate the law that in any process baryon number is conserved.

The stability of the proton can be attributed to this law.

Since it is the lightest of all baryons, the hypothetical products of its decay would have to be non-baryons.

Thus, the decay would violate the conservation of baryon number.

Attempts to unite the electromagnetic, weak and strong interactions into a grand unified theory or GUT have suggested that protons are in fact unstable and decay into leptons.

However, the average lifetime predicted for the protons is ~ 10^{32}years, which makes their decay hard to detect.

Experimental evidence of proton decay has yet to be found.

The conservation of baryon number has no clear theoretical basis nor has any simple symmetry yet been associated with it.

A comprehensive elucidation of the remaining partially or incompletely conserved quantities -*isospin* and *parity*- is beyond the scope of this class.

We will content ourselves with a brief outline of each.

(iv) The Conservation of Isospin

The observation that the strong interaction does not distinguish between nucleons -it treats protons and neutrons identically- underlies the concept of isospin.

Instead of regarding protons and neutrons as totally different species, as far as strong interactions are concerned, they are regarded as being different isospin states of the same underlying nucleon particle.

In a similar way, the three pions, $\pi+$, $\pi-$ and π_0, can be regarded as three different states of the same particle when only strong interactions are involved.

Isospin is mathematically similar to spin, although it has nothing whatsoever to do with angular momentum.

Thus, isospin can be regarded as a vector, usually designated by the greek letter T (tau), in an abstract three-dimensional space, whose magnitude, T, and third component, T_3 or T_z, obey quantization conditions similar to those of spin.

The proton and neutron correspond to different orientations of the isospin vector: the proton corresponds to 'a nucleon with isospin up' (T_3 =+1/2);the neutron to 'a nucleon with isospin down' (T_3 = -1/2).

The notion is analogous to the distinction we make between the two particles we designate as 'an electron with spin up' and 'an electron with spin down'.

Experiments indicate that both T and T_3 are conserved in processes that involve strong interactions.

In processes involving weak interactions neither quantities are conserved.

In those where strong and electromagnetic interactions are involved, only T_3 is conserved. Thus isospin conservation is not absolute.

The symmetry associated with the conservation of T is the rotational symmetry of the strong interaction in the three-dimensional abstract space.

The conservation of T_3 is associated with the symmetry operation called charge reflection -changing the sign of all the electric charges, i.e., replacing +q by -q, and vice versa-under which strong and electromagnetic interactions remain invariant.

If the only interaction we were concerned with was the strong interaction, the symmetry of neutrons and protons would be complete; neutrons could be substituted for protons, and vice versa, without any change occurring in the system.

At the very close distances met within the nucleus this is true to a very good approximation.

However, as the distance between the particles increases, and the effect of the electromagnetic interaction between the protons becomes dominant, the symmetry between the neutrons and protons breaks down.

The isospin symmetry is said to be broken by the electromagnetic interaction.

The rationale that underlies the unification of the weak and electromagnetic interactions into the single electroweak interaction also hinges on the notion that apparently distinct particles, the regular photon and the $W^{\pm}$ and Z^0 bosons, can be regarded as differing forms of the same underlying quantum entity.

However, their initial symmetric configuration is unstable and breaks down spontaneously, producing the asymmetry represented by the different properties they manifest.

Restoring the symmetry requires the investment of energy.

It is estimated that in excess of lTeV ($l0^{12}$eV) is required for the symmetry of these particles to be made explicit.

Such energies are available in terrestrial laboratories to show this and could have been available immediately after the Big Bang, before the initial symmetry was broken.

(v) The Conservation of Parity

In physical terms, parity refers to the operation of space reflection, symbolized by the letter P.

Parity is conserved if the mirror image of a process is itself one which can occur in Nature.

In quantum mechanics, the concept arises from considerations of the relation between the wave functions at the symmetric points, A and A', in a system such as an harmonic oscillator (Fig 6.46a).

Assuming the wave functions are real

$$\psi_A = \pm\psi_{A'} \tag{6.76}$$

When $\psi_A = \psi_{A'}$, we say that the wave function has even parity or P = +l; when $\psi_A = -\psi_{A'}$, the wave function has odd parity or P = -1.

The strong and electromagnetic interactions both conserve parity.

However, as we saw in the context of the decay of radioactive nuclei, it is not conserved by the weak interaction.

Another transformation that is not conserved by the weak interaction is *charge conjugation*, usually symbolized by the letter C, under which all the particles in a system are replaced by their corresponding antiparticles, without making any other changes.

The link between the parity conservation and charge conjugation is illustrated by the decay of the π+ pion into the muon, μ+, and a neutrino, υ, and its charge conjugate decay process:

$$\begin{array}{lccccc} & \pi^+ & \rightarrow & \mu^+ & + & \nu \\ \textbf{Conjugate} & \pi^- & \rightarrow & \mu^- & + & \bar{\nu} \end{array} \qquad (6.77)$$

As Fig 6.50 shows, the charge conjugate decay would require the antineutrino to have negative helicity, which is impossible; antineutrinos always have positive helicity.

Thus, the conjugate decay cannot occur.

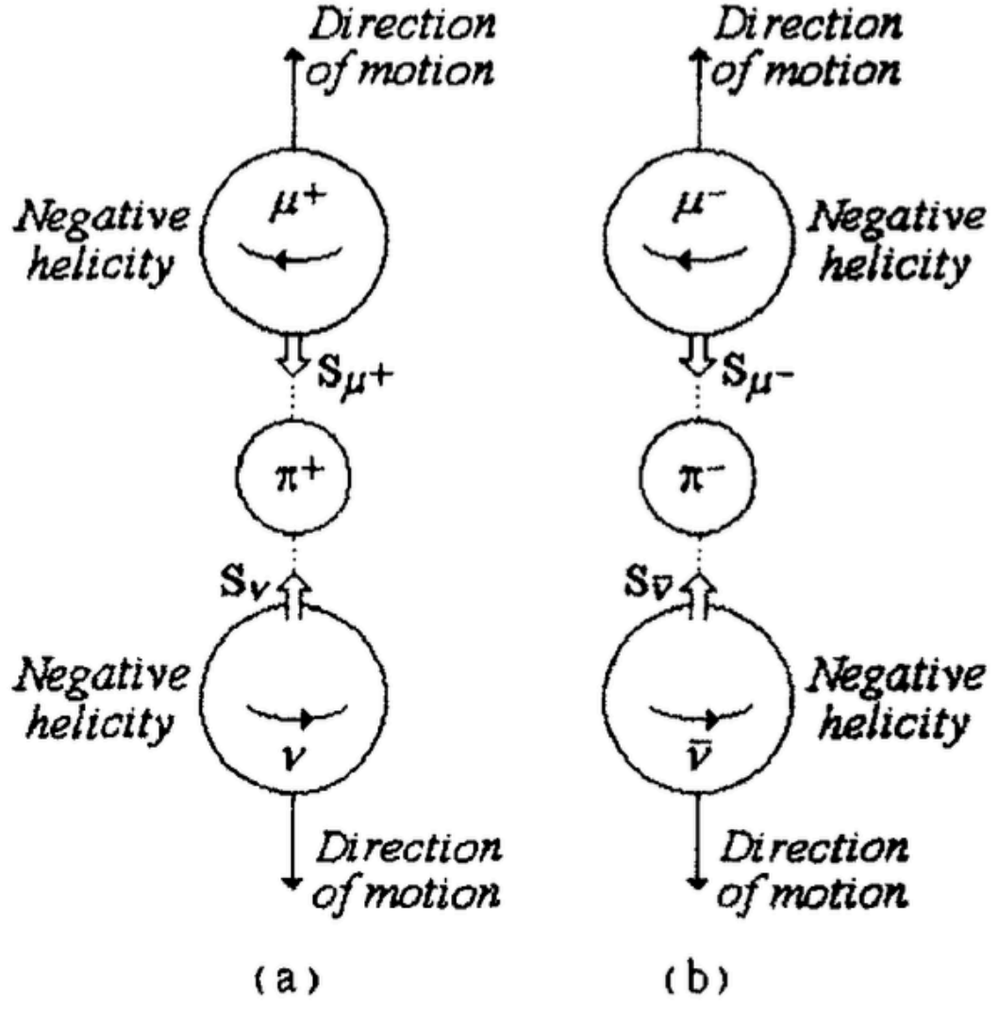


Fig 6.50 The decay of the π+ pion and its charge conjugate. The charge conjugate decay of the π- pion cannot occur as shown, as this would require that the antineutrino possess a negative helicity.

However, suppose that the π+ pion decay undergoes two symmetry operations, the P (parity) operation followed by the C (charge conjugate) operation, resulting in the combined operation CP. The π- decay process that this gives can occur, for it assigns a positive helicity to the antineutrino (Fig 6.51).

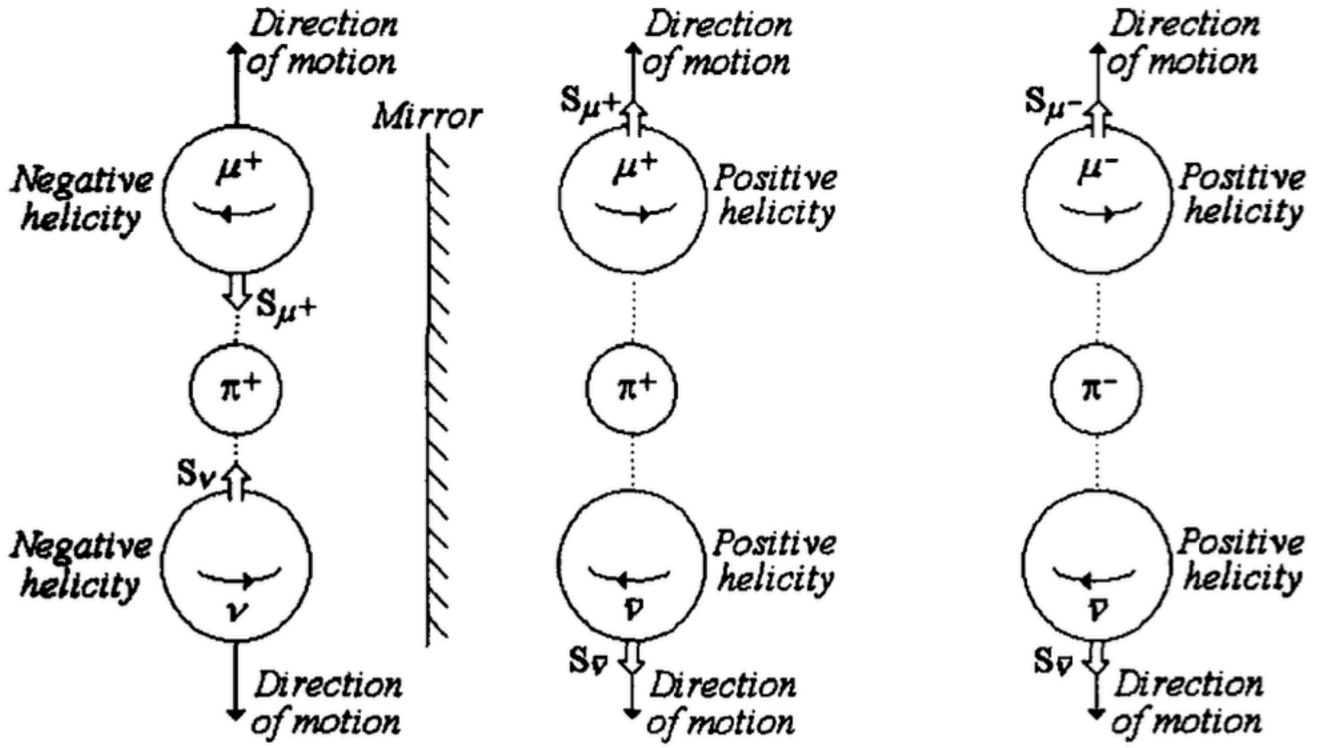


Fig 6.51 The CP operation applied to π+ decay. Two successive symmetry operations, are applied to the π+ decay. First, the P operation (reflection in a plane mirror) which changes the helicity of the particles, followed by the C (charge conjugation) which replaces each particle by its antiparticle. The π- decay process this gives can occur, for it assigns a positive helicity to the antineutrino.

Although, in general, the weak interaction remains invariant under CP operations, an instance in which it does not conserve CP has been observed experimentally.

At this stage, you might well ask whether there is any symmetry operation under which all the known physical processes are invariably invariant.

There is good reason to believe that all physical laws are invariant under the combined symmetry operations of time reversal, T, parity, P, and charge conjugation, C, known as the CPT operation.

The implication of this is that a violation of one of the parameters must be accompanied by a compensating violation of another parameter.

If CP is violated, the time parameter must also be violated so that CPT may be conserved.

Such a process would not be time symmetric, i.e., it could only occur in one time direction.

Though we are familiar with this asymmetry in macroscopic systems, in which the second law of thermodynamics holds sway, the import of its occurrence at the particle level is not fully understood.

If indeed there is an intrinsic asymmetry in Nature, the observation that the universe is composed almost exclusively of matter, with a very small antimatter component could be explained.

For if the symmetry between matter and antimatter were perfect, the material universe could not exist.

Composed equally of matter and antimatter, which would interact and annihilate one another, it would vanish in a titanic burst of radiation.

It is, perhaps, so that we could exist that a slight imperfection was allowed for at the Big Bang.

A More Up-To-Date Addendum on Elementary Particles to finish up!

THE FUNDAMENTAL FORCES IN NATURE

To understand the properties of elementary particles, we must be able to describe the forces between them.

Particles in nature are subject to four fundamental forces; in order of decreasing strength, these are the strong force, the electromagnetic force, the weak force, and the gravitational force.

The **strong force** is responsible for binding quarks tightly together to form protons, neutrons, and other heavy particles.

It is extremely short-range and is negligible for separations greater than approximately 10^{-15} m.

The nuclear force that binds neutrons and protons in nuclei is currently believed to be a residual effect of the more basic strong force between quarks, much as the molecular force binding electrically neutral atoms together in molecules is a residual electrical interaction.

The **electromagnetic force**, which binds electrons and protons within atoms and molecules to form ordinary matter, is approximately two orders of magnitude weaker than the strong force.

It is a long-range force that decreases in strength as the inverse square of the separation between interacting particles.

The **weak force** is a short-range force that accounts for the beta decay of nuclei and the decay of heavier quarks and leptons.

Its strength is only about 10^{-6} times that of the strong force.

As we shall discuss later, scientists now believe that the weak and electromagnetic forces are two manifestations of a unified force called the electroweak force.

Finally, the **gravitational force** is a long-range force that has a strength only about 10^{-43} times that of the strong force.

Although this familiar interaction holds the planets, stars, and galaxies together, its effect on elementary particles is negligible.

Classically, the entity that is responsible for transmitting a force from one particle to another is the field.

For example, a positive electric charge produces an electric field in space, which in turn exerts an attractive force on a nearby negative charge.

Furthermore, the field can carry energy and momentum from one particle to the other.

According to quantum field theories, the energy and momentum of all fields are quantized, and the quantum that carries a "chunk" of momentum and energy from one type of particle to another is called a **field particle**.

In particle physics, interactions between particles are described in terms of the exchange of field particles, or quanta, *which are all bosons*.

In the case of the familiar electromagnetic interaction, for instance, the field particles are photons.

In the language of modern physics, one can say that the electromagnetic force is *mediated*, or carried, by photons and that photons are the quanta of the electromagnetic field.

Likewise, the strong force is mediated by field particles called *gluons*, the weak force is mediated by particles called the $W^{\pm}$ and Z^0 *bosons*, and the gravitational force is carried by quanta of the gravitational field called gravitons.

These interactions, the particles they act on, their ranges, their relative strengths, and the corresponding field particles, are summarized in Table 15.1.

Note that the fourth column of Table 15.1 presents another way of classifying interactions—by means of the observed lifetime of a decaying particle.

In fact, with only a few exceptions, the lifetimes of decaying particles are excellent indicators of what interaction has caused the decay, with shorter lifetimes being associated with stronger forces.

As shown in the Table, particles decaying via the strong force are usually the shortest-lived; next come those decaying via electromagnetic forces; and finally, particles decaying by the weak interaction have the longest observed lifetimes.

Table Particle Interactions

Interaction (Force)	Particles Acted on by Force	Relative Strength[a]	Typical Lifetimes for Decays via a Given Interaction	Range of Force	Force-Carrying Particle Exchanged
Strong	Quarks, hadrons	1	$\leq 10^{-20}$ s	Short ($\approx$1 fm)	Gluon
Electromagnetic	Charged particles	$\approx 10^{-2}$	$\approx 10^{-16}$ s	Long (∞)	Photon
Weak	Quarks, leptons	$\approx 10^{-6}$	$\geq 10^{-10}$ s	Short ($\approx 10^{-3}$ fm)	$W^{\pm}$, Z^0 bosons
Gravitational	All particles	$\approx 10^{-43}$	?	Long (∞)	Graviton[b]

[a]For two u quarks at 3×10^{-17} m.
[b]Not experimentally detected.

POSITRONS AND OTHER ANTIPARTICLES

In the 1920s the English theoretical physicist Paul Adrien Maurice Dirac(1902–1984) developed a version of quantum mechanics that incorporated special relativity.

Dirac's theory automatically explained the origin of the electron's spin and its magnetic moment.

However, it also presented a major difficulty. Dirac's relativistic wave equation required solutions corresponding to both positive and negative energies for free electrons.

We can easily see this from the expression for the total relativistic energy of an electron, $E = \pm\sqrt{p^2c^2 + m_e^2c^4}$, which one would expect an electron in a state of positive energy to make a rapid transition to one of these lower energy states, emitting a photon in the process.

Eventually all electrons in the universe would end up locked in negative energy states, contradicting the common observation of electrons with positive total energies.

Dirac avoided this difficulty by postulating that all negative energy states are filled.

Electrons that occupy the negative energy states are called collectively the "Dirac sea."

Electrons in the Dirac sea are not directly observable, because the Pauli exclusion principle does not allow them to react to external forces— there are no available states to which an electron can make a transition in response to an external force.

Therefore, an electron in such a state acts as an isolated system, unless enough energy ($\geq 2m_ec^2$) is supplied to excite the electron to a positive energy state.

Such an excitation causes one of the negative energy states to become vacant, as in Figure, leaving a hole in the sea of filled states.

The hole can react to external forces and is observable.

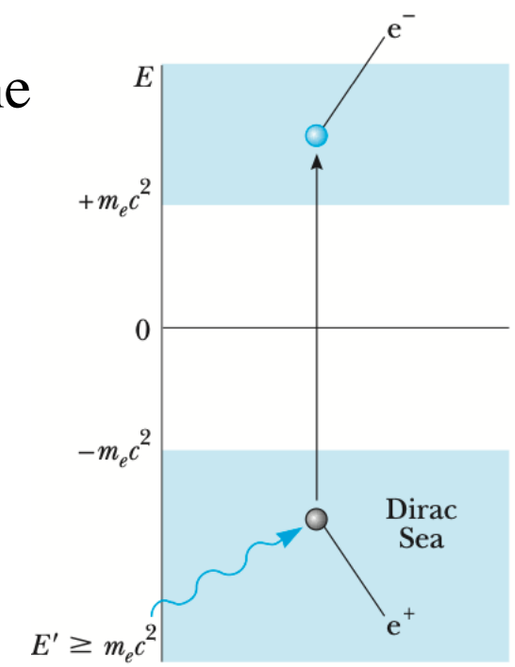


Figure: Dirac's model for the existence of antielectrons (positrons). The states lower in energy than m_ec^2 are filled with electrons (the Dirac sea). One of these electrons can make a transition out of its state only if it is provided with energy equal to or larger than $2m_ec^2$. This leaves a vacancy in the Dirac sea, which can behave as a particle identical to the electron except for its positive charge.

The hole reacts in a way similar to that of the electron, except that it has a positive charge—it is the **antiparticle** to the electron.

The electron's antiparticle, the **positron**, has a rest energy of 0.511 MeV and a positive charge of 1.60 x 10^{-19} C.

Carl Anderson (1905–1991) observed the positron experimentally in 1932, and in 1936 he was awarded a Nobel prize for his achievement.

Anderson discovered the positron while examining tracks created in a cloud chamber by electron-like particles of positive charge.

These early experiments used cosmic rays—mostly energetic protons passing through interstellar space—to initiate high-energy reactions on the order of several GeV.

To discriminate between positive and negative charges, Anderson placed the cloud chamber in a magnetic field, causing moving charges to follow curved paths.

He noted that some of the electron-like tracks deflected in a direction corresponding to a positively charged particle.

Since Anderson's discovery, positrons have been observed in a number of experiments.

A common source of positrons is pair production.

In this process, a gamma-ray photon with sufficiently high energy interacts with a nucleus, and an electron–positron pair is created from the photon.

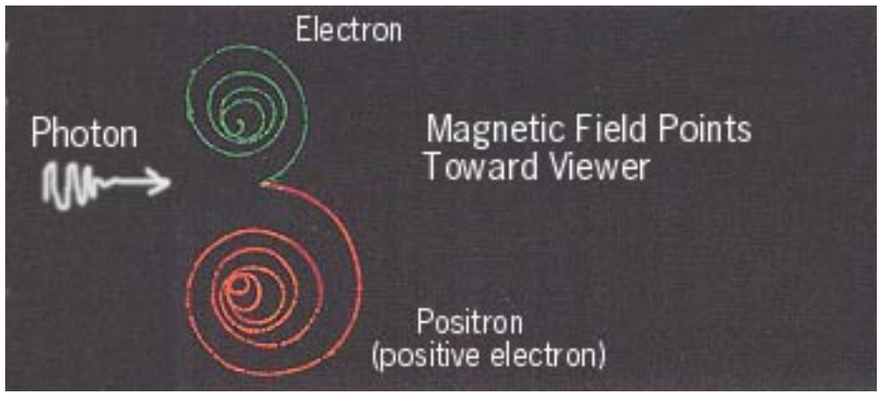
The presence of the nucleus allows the principle of conservation of momentum to be satisfied.

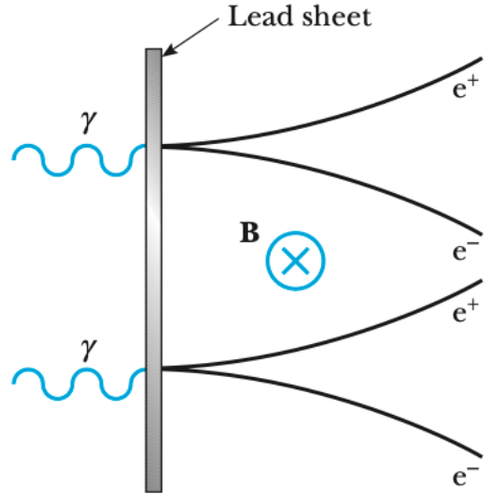
Because the total rest energy of the electron–positron pair is $2m_ec^2 = 1.02$ MeV, the photon must have at least this much energy to create an electron–positron pair.

Therefore, the electromagnetic energy of a photon is converted mainly to rest energy of the electron and positron in accordance with Einstein's relationship $E = mc^2$.

If the gamma-ray photon has energy in excess of the rest energy of the electron–positron pair, the excess appears as kinetic energy of the two particles.

Figure shows tracks of electron–positron pairs created by 300-MeV gamma rays striking a lead sheet.





A sketch of the pertinent pair-production events. Note that the positrons deflect upward while the electrons deflect downward in an applied magnetic field directed into the page.

The reverse process can also occur.

Under the proper conditions, an electron and a positron can annihilate each other to produce two gamma-ray photons with a combined energy of at least 1.02 MeV:

$$e^- + e^+ \longrightarrow 2\gamma$$

A practical application of electron–positron annihilation occurs in the medical diagnostic technique called **positron emission tomography (PET)**.

The patient is injected with a glucose solution containing a radioactive substance that decays by positron emission, and the material is carried by the blood throughout the body.

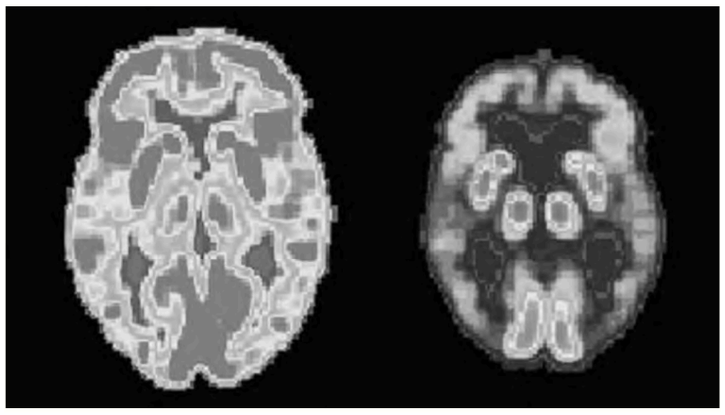
A positron emitted during a decay event in one of the radioactive nuclei in the glucose solution annihilates with an electron in the surrounding tissue, resulting in two gamma-ray photons emitted in opposite directions.

A gamma detector surrounding the patient pinpoints the source of the photons and, with the assistance of a computer, displays an image of the sites at which the glucose accumulates.

Glucose is metabolized rapidly in cancerous tumors and accumulates at those sites, providing a strong signal for a PET detector system.

The images from a PET scan can indicate a wide variety of disorders in the brain, including Alzheimer's disease (Figure).

In addition, because glucose metabolizes more rapidly in active areas of the brain, a PET scan can indicate which areas of the brain are involved in activities in which a patient is engaging at the time of the scan, such as language use, music, and vision.



PET scans of the brain of a healthy older person (left) and that of a patient suffering from Alzheimer's disease (right).

Prior to 1955, on the basis of the Dirac theory, it was expected that every particle had a corresponding antiparticle, but antiparticles such as the antiproton and antineutron had not been detected experimentally.

Since the relativistic Dirac theory had had some failures (it predicted the wrong-size magnetic moment for the proton) as well as many successes, it was important to determine whether the antiproton really existed.

In 1955 a team led by Emilio Segrè (1905 – 1989, Italian-American physicist) and Owen Chamberlain (b. 1920, American physicist) used the Bevatron particle accelerator at the University of California, Berkeley, to produce both antiprotons and antineutrons.

They thus established with certainty the existence of antiparticles.

For this work Segrè and Chamberlain received the Nobel prize in 1959.

It is now accepted that every particle has a corresponding antiparticle of equal mass and spin and of equal and opposite charge, magnetic moment, and strangeness.

The property of strangeness is explained later.

The only exceptions to these rules for particles and antiparticles are the neutral photon, pion, and eta, each of which is its own antiparticle.

MESONS AND THE BEGINNING OF PARTICLE PHYSICS

In the mid-1930s, physicists had a fairly simple view of the structure of matter.

The building blocks were the proton, the electron, and the neutron.

Three other particles were known or had been postulated at the time: the photon, the neutrino, and the positron.

These six particles were considered the fundamental constituents of matter.

With this marvelously simple picture of the world, however, no one could provide an answer to an important question:

Since the many protons close together in a nucleus should strongly repel each other because of their like charges, what is the nature of the force that holds the nucleus together?

Scientists recognized that this mysterious force must be much stronger than anything encountered in nature up to that time.

In 1935 the Japanese physicist Hideki Yukawa (1907–1981) proposed the first theory to successfully explain the nature of the nuclear force, an effort that later earned him the Nobel prize.

To understand Yukawa's theory, it is useful to recall that in the modern view of electromagnetic interactions, *charged particles interact by exchanging photons*.

Yukawa used this idea to explain the nuclear force by proposing a new particle whose exchange between nucleons in the nucleus produces the nuclear force.

Furthermore, he established that the range of the force is inversely proportional to the mass of this particle and predicted that the mass would be about 200 times the mass of the electron.

Since the new particle would have a mass between that of the electron and that of the proton, it was called a **meson** (from the Greek *meso*, "middle").

In an effort to substantiate Yukawa's predictions, physicists began an experimental search for the meson by studying cosmic rays entering the Earth's atmosphere.

In 1937 Carl Anderson and his collaborators discovered a particle of mass 106 MeV/c2, about 207 times the mass of the electron.

However, subsequent experiments showed that the particle interacted very weakly with matter and hence could not be the carrier of the strong force.

The puzzling situation inspired several theoreticians to propose that there were two mesons with slightly different masses.

This idea was confirmed by the discovery in 1947 of the pi (π) meson, or simply **pion**, by Cecil Frank Powell (1903–1969) and Giuseppe P. S. Occhialini (b. 1907).

The particle discovered by Anderson in 1937, the one thought to be a meson, is not really a meson.

Instead, it takes part in the weak and electromagnetic interactions only, and it is now called the **muon**.

The pion, Yukawa's carrier of the nuclear force, comes in three varieties, corresponding to three charge states: π^+, π^- and π^0.

The π^+ and π^- particles have masses of 139.6 MeV/c^2, and the π^0 has a mass of 135.0 MeV/c^2.

Both pions and muons are unstable particles.

For example, the π^-, which has a mean lifetime of 2.6 x10^{-8} s, first decays to a muon and a muon antineutrino.

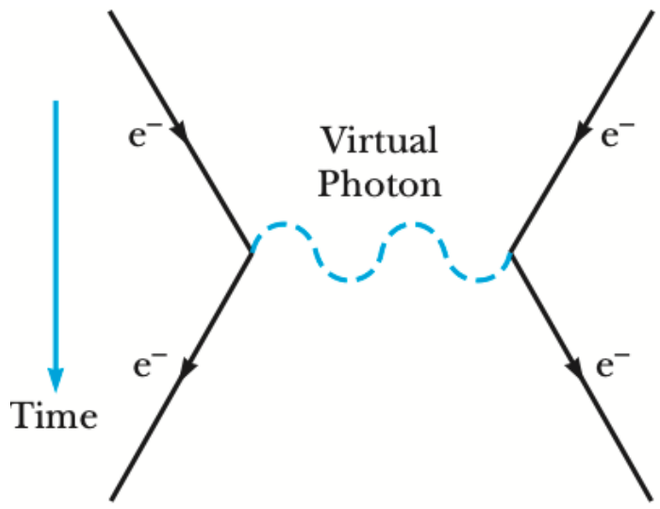
The muon, which has a mean lifetime of 2.2 μs, then decays to an electron, a neutrino, and an electron antineutrino:

$$\pi^- \longrightarrow \mu^- + \bar{\nu}_\mu$$

$$\mu^- \longrightarrow e^- + \nu_\mu + \bar{\nu}_e$$

The interaction between two particles can be represented in a useful diagram called a **Feynman diagram**, developed by the American physicist Richard P. Feynman (1918–1988).

Figure is such a diagram for the electromagnetic interaction between two electrons.



Feynman diagram showing how a photon carries the electromagnetic force between two interacting electrons. The blue arrow shows the direction of increasing time.

A Feynman diagram is a qualitative graph of time on the vertical axis versus space on the horizontal axis.

It is qualitative in the sense that actual values of time and space are not important, but the overall appearance of the graph conveniently serves to categorize different interaction processes.

In this simple case, a photon is the field particle that mediates the electromagnetic force between the electrons.

The photon transfers energy and momentum from one electron to the other in this interaction.

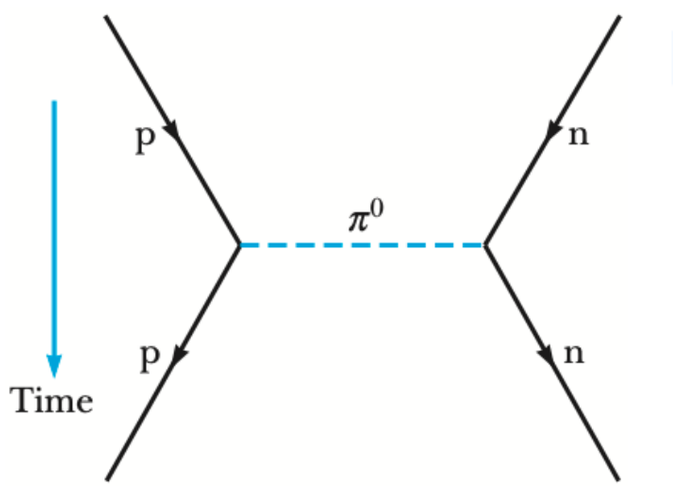
The photon is called a **virtual photon** because it is emitted and reabsorbed without being detected.

Virtual photons can violate the law of conservation of energy by E for a very short time, t, provided that $\Delta E\, \Delta t \approx \hbar/2$ (from the minimum form of the uncertainty principle).

Note that such quantum violations of energy conservation take place only in the short term and that system energy is conserved in the long run when the photon is reabsorbed by the other electron.

Now consider the pion exchange between a proton and a neutron that transmits the nuclear force according to Yukawa (Figure).

Feynman diagram representing a proton interacting with a neutron via the strong force. In this case, the pion mediates the strong force. The blue arrow shows the direction of increasing time.



We can reason that the energy ΔE needed to create a pion of mass m_π is given by Einstein's equation $\Delta E = m_\pi c^2$.

Again, the very existence of the pion would violate conservation of energy if it lasted for a time greater than $\Delta t \approx \hbar/2\,\Delta E$, where ΔE is the energy of the pion and Δt is the time it takes the pion to travel from one nucleon to the other.

Therefore,

$$\Delta t \approx \frac{\hbar}{2\,\Delta E} = \frac{\hbar}{2m_\pi c^2}$$

Because the pion cannot travel faster than the speed of light, the maximum distance d it can travel in a time Δt is $c\Delta t$.

Using Equation 15.2 and $d = c\Delta t$, we find this maximum distance to be

$$d \approx \frac{\hbar}{2m_\pi c}$$

We know that the range of the nuclear force is approximately 1.0×10^{-15} m.

Using this value for d, we calculate the rest energy of the pion to be

$$\begin{aligned} m_\pi c^2 &\approx \frac{\hbar c}{2d} = \frac{197\ \text{MeV}\cdot\text{fm}}{2(1 \times 10^{-15}\ \text{m})} \\ &\approx 100\ \text{MeV} \end{aligned}$$

This corresponds to a mass of 100 MeV/c^2 (approximately 250 times the mass of the electron), a value in reasonable agreement with the observed pion mass.

The concept we have just described is quite important.

In effect, it says that a proton can change into a proton plus a pion as long as it returns to its original state in a very short time.

Physicists often say that a nucleon undergoes "fluctuations" as it emits and absorbs pions.

As we have seen, these fluctuations are a consequence of a combination of quantum mechanics (through the uncertainty principle) and special relativity (through Einstein's energy–mass relationship, $E = mc^2$).

Digression about Richard Feynman

Richard Phillips Feynman was a brilliant theoretical physicist who shared the 1965 Nobel Prize in Physics with Julian S.Schwinger and Shinichiro Tomonaga for their fundamental work in the principles of quantum electrodynamics.

Feynman's many important contributions to physics include the invention of simple diagrams to represent particle interactions graphically, the theory of the weak interaction of subatomic particles, a reformulation of quantum mechanics, the theory of superfluid helium, and his contribution to physics education through the magnificent three-volume text *The Feynman Lectures on Physics*.

Feynman did his undergraduate work at MIT and received his Ph.D. in 1942 from Princeton University, where he studied under John Archibald Wheeler.

During World War II he worked on the Manhattan Project at Princeton and at Los Alamos, New Mexico.

He then joined the faculty at Cornell University in 1945 and in 1950 was appointed professor of physics at California Institute of Technology, where he remained for the rest of his career.

It is well known that Feynman had a passion for finding new and better ways to formulate each problem or, as he would say, "turning it around."

In the early part of his career, he was fascinated with electrodynamics and developed an intuitive view of quantum electrodynamics.

Convinced that the electron could not interact with its own field, he said, "That was the beginning, and the idea seemed so obvious to me that I fell deeply in love with it. . . ." Often called the outstanding intuitionist of our age, he said in his Nobel acceptance speech, "Often, even in a physicist's sense, I did not have a demonstration of how to get all of these rules and equations, from conventional electrodynamics.. . . I never really sat down, like Euclid did for the geometers of Greece, and made sure that you could get it all from a single set of axioms."

In 1986 Feynman was a member of the presidential commission to investigate the explosion of the space shuttle Challenger.

In this capacity, he performed a simple experiment for the commission members that showed that one of the shuttle's O-ring seals was the likely cause of the disaster.

After placing a seal in a pitcher of ice water and squeezing it with a clamp, he demonstrated that the seal failed to spring back into shape once the clamp was removed.

Feynman worked in physics with a style commensurate with his personality—that is, with energy, vitality, and humor.

The following quotations from some of his colleagues hint at the great impact he made on the scientific community.

Murray Gell-Mann: "A brilliant, vital, and amusing neighbor, Feynman was a stimulating (if sometimes exasperating) partner in discussions of profound issues—we would exchange ideas and silly jokes in between bouts of mathematical calculation—we struck sparks off each other, and it was exhilarating."

David Pines: "Reading Feynman is a joy and a delight, for in his papers, as in his talks, Feynman communicated very directly, as though the reader were watching him derive the results at the blackboard."

David L. Goodstein: "He loved puzzles and games. In fact, he saw all the world as a sort of game, whose progress of 'behavior' follows certain rules, some known, some unknown. . . . Find places or circumstances where the rules don't work, and invent new rules that do."

Valentine L. Telegdi: "Feynman was not a theorist's theorist, but a physicist's physicist and a teacher's teacher."

Laurie M. Brown, one of his graduate students at Cornell, noted that Feynman, a playful showman, was "undervalued at first because of his rough manners [but] in the end triumphs through native cleverness, psychological insight, common sense and the famous Feynman humor. . . . Whatever else Dick Feynman may have joked about, his love for physics approached reverence."

This first part of the discussion has dealt with early ideas about particles that carry the nuclear force, namely the pions, and the mediators of the electromagnetic force, photons.

Current ideas indicate that the nuclear force is more accurately described as an average or residual effect of the strong color force between quarks, as will be explained shortly.

The graviton, which is the mediator of the gravitational force, has yet to be observed.

The $W^{\pm}$ and Z^0 particles that mediate the weak force were discovered in 1983 by the Italian physicist Carlo Rubbia (b. 1934) and his associates, using a proton – antiproton collider.

Rubbia and Simon van der Meer, both at CERN near Geneva, Switzerland, shared the 1984 Nobel prize for the discovery of the $W^{\pm}$ and Z^0 particles and the development of the proton – antiproton collider.

In this accelerator, protons and antiprotons that have a momentum of 270 GeV/c undergo head-on collisions with each other.

In some of the collisions $W^{\pm}$ and Z^0 particles are produced, which in turn are identified by their decay products.

CLASSIFICATION OF PARTICLES

All particles other than field particles can be classified into two broad categories, hadrons and leptons, according to their interactions.

Hadrons

Particles that interact through the strong force are called **hadron**s.

The two classes of hadrons, **mesons** and **baryons**, are distinguished by their masses and spins.

Mesons all have spin 0 or 1, with masses between that of the electron and that of the proton.

All mesons are known to decay finally into electrons, positrons, neutrinos, and photons.

The pion is the lightest of known mesons; it has a mass of approximately 140 MeV/c^2 and a spin of 0.

Another is the K meson, with a mass of approximately 500 MeV/c^2 and a spin of 0.

Baryons, the second class of hadrons, have masses equal to or greater than the proton mass (baryon means "heavy" in Greek), and their spins are always odd half-integer values ($\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$ …., etc.).

Protons and neutrons are baryons, as are many other particles.

With the exception of the proton, all baryons decay in such a way that the end products include a proton.

For example, the baryon called the Ξ^- hyperon first decays to the Λ^0 baryon and a π- in about 10^{-10} s.

The Λ^0 then decays to a proton and a π- in approximately 3 x10^{-10} s.

It is important to note that hadrons are composite particles, not point particles, and have a measurable size of about 1 fm (10^{-15} m).

Hadrons are composed of more elemental units called quarks, which are believed to be truly structureless point particles.

Mesons consist of two quarks and baryons of three.

For now, however, we defer discussion of the ultimate constituents of hadrons until later and continue with our empirical classification of particles.

Table lists important properties of the leptons and some hadrons.

The symbols B, Le, L_μ , L_τ , and S stand for baryon, electron, muon, and tau numbers and strangeness, respectively, and are explained later.

Leptons

Leptons (from the Greek *leptos*, meaning "small" or "light") are a group of particles that participate in the electromagnetic and weak interactions.

All leptons have spins of 1/2.

Unlike hadrons, which have size and structure, **leptons appear to be truly elementary point-like particles with no structure**.

Also unlike hadrons, the number of known leptons is small.

Table Some Particles and Their Properties

Category	Particle Name	Symbol	Anti-particle	Mass (MeV/c^2)	B	L_e	L_μ	L_τ	S	Lifetime (s)	Principal Decay Modes[a]
Leptons	Electron	e^-	e^+	0.511	0	+1	0	0	0	Stable	
	Electron-neutrino	ν_e	$\overline{\nu}_e$	$<2.8 \times 10^{-6}$	0	+1	0	0	0	Stable	
	Muon	μ^-	μ^+	105.7	0	0	+1	0	0	2.19×10^{-6}	$e^- \overline{\nu}_e \nu_\mu$
	Muon-neutrino	ν_μ	$\overline{\nu}_\mu$	$<3.5 \times 10^{-6}$	0	0	+1	0	0	Stable	
	Tau	τ^-	τ^+	1784	0	0	0	+1	0	3.3×10^{-13}	$\mu^- \overline{\nu}_\mu \nu_\tau$, $e^- \overline{\nu}_e \nu_\tau$
	Tau-neutrino	ν_τ	$\overline{\nu}_\tau$	$<8.4 \times 10^{-6}$	0	0	0	+1	0	Stable	
Hadrons											
Mesons	Pion	π^+	π^-	139.6	0	0	0	0	0	2.60×10^{-8}	$\mu^+ \nu_\mu$
		π^0	Self	135.0	0	0	0	0	0	0.83×10^{-16}	2γ
	Kaon	K^+	K^-	493.7	0	0	0	0	+1	1.24×10^{-8}	$\mu^+ \nu_\mu$ $\pi^+ \pi^0$
		K_S^0	$\overline{K}_S^0$	497.7	0	0	0	0	+1	0.89×10^{-10}	$\pi^+ \pi^-$ $2\pi^0$
		K_L^0	$\overline{K}_L^0$	497.7	0	0	0	0	+1	5.2×10^{-8}	$\pi^\pm e^\mp \overline{\nu}_e$ $3\pi^0$ $\pi^\pm \mu^\mp \overline{\nu}_\mu$
	Eta	η	Self	548.8	0	0	0	0	0	$<10^{-18}$	2γ, $3\pi^0$
		η'	Self	958	0	0	0	0	0	2.2×10^{-21}	$\eta \pi^+ \pi^-$
Baryons	Proton	p	$\overline{p}$	938.3	+1	0	0	0	0	Stable	
	Neutron	n	$\overline{n}$	939.6	+1	0	0	0	0	624	$p e^- \overline{\nu}_e$
	Lambda	Λ^0	$\overline{\Lambda}^0$	1115.6	+1	0	0	0	−1	2.6×10^{-10}	$p\pi^-$, $n\pi^0$
	Sigma	Σ^+	$\overline{\Sigma}^-$	1189.4	+1	0	0	0	−1	0.80×10^{-10}	$p\pi^0$, $n\pi^+$
		Σ^0	$\overline{\Sigma}^0$	1192.5	+1	0	0	0	−1	6×10^{-20}	$\Lambda^0 \gamma$
		Σ^-	$\overline{\Sigma}^+$	1197.3	+1	0	0	0	−1	1.5×10^{-10}	$n\pi^-$
	Delta	Δ^{++}	$\overline{\Delta}$	1230	+1	0	0	0	0	6×10^{-24}	$p\pi^+$
		Δ^+	$\overline{\Delta}^-$	1231	+1	0	0	0	0	6×10^{-24}	$p\pi^0$, $n\pi^+$
		Δ^0	$\overline{\Delta}^0$	1232	+1	0	0	0	0	6×10^{-24}	$n\pi^0$, $p\pi^-$
		Δ^-	$\overline{\Delta}^+$	1234	+1	0	0	0	0	6×10^{-24}	$n\pi^-$
	Xi	Ξ^0	$\overline{\Xi}^0$	1315	+1	0	0	0	−2	2.9×10^{-10}	$\Lambda^0 \pi^0$
		Ξ^-	$\overline{\Xi}^+$	1321	+1	0	0	0	−2	1.64×10^{-10}	$\Lambda^0 \pi^-$
	Omega	Ω^-	$\overline{\Omega}^+$	1672	+1	0	0	0	−3	0.82×10^{-10}	$\Xi^- \pi^0$, $\Xi^0 \pi^-$, $\Lambda^0 K^-$

[a]Notations in this column such as $p\pi^-$, $n\pi^0$ indicate two possible decay modes. In this case, the two possible decays are $\Lambda^0 \rightarrow p + \pi^-$ and $\Lambda^0 \rightarrow n + \pi^0$.

Currently, scientists believe there are only six leptons: the electron, the muon, the tau, and — associated with each of these particles — three different neutrinos, the electron neutrino (ν_e), the muon neutrino (ν_μ), and the tau neutrino (ν_τ).

We now classify the six known leptons into three pairs called families:

$$\begin{pmatrix} e^- \\ \nu_e \end{pmatrix} \quad \begin{pmatrix} \mu^- \\ \nu_\mu \end{pmatrix} \quad \begin{pmatrix} \tau^- \\ \nu_\tau \end{pmatrix}$$

The table shows that the tau lepton is actually quite massive and has a mass about twice that of the proton.

Also note in the table that each lepton has an antiparticle.

It is quite unexpected and interesting that neutrinos and antineutrinos each have a distinct helicity, or relation between linear momentum and spin directions.

The spin of a neutrino is opposite to its direction of travel, and the spin of an antineutrino is parallel to its direction of travel.

Current evidence suggests that neutrinos have a small mass of several eV/c^2 as shown in the table.

An interesting time-of-flight method is able to set an upper limit on neutrino mass.

As we shall see, a firm knowledge of the neutrino's mass has great significance in physical models of energy production in stars and in grand unified theories of elementary particles.

The Solar Neutrino Mystery and Neutrino Oscillations

Conclusive measurements of the ability of neutrinos to change from one type or "flavor" to another, along with indirect measurements of neutrino masses have recently been made at the Sudbury Neutrino Observatory (SNO) in Canada and the Super Kamiokanda detector in Japan.

These findings solve a puzzle over 40 years old and provide new confidence that physicists really do understand how energy is produced in the Sun's core.

The warming sunlight we receive on Earth should be accompanied by billions of neutrinos per square centimeter per second.

In particular, electron neutrinos are produced in the Sun's fusion engine by typical reactions like

$$\mathrm{p} + \mathrm{p} \longrightarrow {}^{2}_{1}\mathrm{H} + \mathrm{e}^{+} + \nu_{\mathrm{e}} + 0.42\ \mathrm{MeV}$$

and the boron decay

$${}^{8}_{5}\mathrm{B} \longrightarrow {}^{8}_{4}\mathrm{Be} + \mathrm{e}^{+} + \nu_{\mathrm{e}} + 14.6\ \mathrm{MeV}$$

Measurements of the e flux dating back to the 1960s have, however, been consistently mysterious because *only about one-third of the expected flux has been observed*.

The expected flux of solar electron neutrinos was calculated using the trusted standard solar model of the Sun's properties and energy production.

One proposed explanation of the mystery is that some ν_{e}'s have changed into ν_{μ}'s or ν_{τ}'s during the journey from the Sun to the Earth, and these ν_{μ}'s or ν_{τ}'s would not show up in detectors designed to spot ν_{e}'s.

The basic idea is that the change from ν_e to ν_μ, say, would not be permanent but would just be part of an ongoing oscillation between ν_e and ν_μ .

The best part is that the frequency of oscillation would depend on the masses.

The measurement of **neutrino oscillation** frequencies could indirectly determine neutrino masses, which are small and extremely hard to measure directly.

We can think of each flavor of neutrino— ν_e, ν_μ, ν_τ — as a linear combination of three other neutrinos ν_1, ν_2, ν_3 each having a different mass m_1, m_2, and m_3 and different energy for a given momentum.

Since the time dependence of each state ν_1, ν_2, ν_3 is oscillatory and of the form exp(-iEt/$\hbar$), with time, the states ν_1, ν_2, ν_3 move out-of-phase with each other and interfere, producing an oscillation between flavors.

The latest findings from SNO scientists, who used a new detection technique equally sensitive to electron, muon, and tau neutrinos, conclusively show that the missing electron neutrinos are accounted for by oscillation of ν_e's into ν_μ's and ν_τ's.

Further, plans are under way at SNO to study how neutrino oscillation depends on the passage of neutrinos through matter.

This can be done by observing day– night differences in neutrino flux from the Sun, because at night neutrinos must penetrate the whole Earth to reach the detector.

CONSERVATION LAWS

Conservation laws are important to an understanding of why certain decays and reactions occur and others do not.

In general, the laws of conservation of energy, linear momentum, angular momentum, and electric charge provide us with a set of rules that all processes must absolutely follow.

For example, conservation of electric charge requires that the total charge before a reaction equals the total charge after the reaction.

Certain new conservation laws are important in the study of elementary particles.

Two of these laws, concerning baryon number and lepton number, are described now, and others will be discussed later.

Although the two described here have no theoretical foundation, they are supported by abundant empirical evidence and serve to indicate probable reactions, but do not absolutely hold 100% of the time.

Baryon Number

To apply conservation of baryon number, we assign a baryon number B = +1 for all baryons, B = -1 for all antibaryons, and B = 0 for all other particles.

The **law of conservation of baryon number** states that *whenever a nuclear reaction or decay occurs, the sum of the baryon numbers before the process must equal the sum of the baryon numbers after the process*.

An equivalent statement is that the net number of baryons remains constant in any process.

If baryon number is absolutely conserved, the proton must be absolutely stable.

If it were not for the law of conservation of baryon number, the proton could decay to a positron and a neutral pion.

However, such a decay has never been observed.

At present, we can say only that the proton has a half-life of at least 10^{32} years (the estimated age of the Universe is only about 10^{10} years).

In one recent version of a grand unified theory, physicists predicted that the proton is unstable.

According to this theory, baryon number is not absolutely or perfectly conserved.

Lepton Number

From observations of commonly occurring decays of the muon and tau we arrive at three conservation laws involving lepton numbers, one for each variety of lepton.

The **law of conservation of electron-lepton number(lepton flavor conservation)** states that *the sum of the electron-lepton numbers* before a reaction or decay must equal the sum of the electron-lepton numbers after the *reaction or decay*.

The electron and the electron neutrino are assigned a positive lepton number, $L_e = +1$; the antileptons e^+ and $\bar{\nu}_e$ are assigned a negative lepton number, $L_e = -1$; all others (the muon and tau families) have $L_e = 0$.

For example, consider the decay of the neutron

$$n \longrightarrow p + e^- + \bar{\nu}_e$$

Before the decay the electron-lepton number is $L_e = 0$; after the decay it is $0 + 1 + (-1) = 0$.

Thus the electron-lepton number is conserved, although not perfectly in view of the proven neutrino oscillation previously mentioned.

It is important to recognize that the baryon number must also be conserved.

We can easily check this by noting that before the decay $B = +1$, and after the decay $B = +1 + 0 + 0 = 1$.

Similarly, when a decay involves muons, the muon-lepton number, L_μ , is conserved.

The μ^- and ν_μ the are assigned positive numbers, $L_\mu = +1$; the antimuons μ^+ and $\bar{\nu}_\mu$ are assigned negative numbers, $L_\mu = -1$ and all others have $L_\mu = 0$.

Finally, the tau-lepton number, L_τ , is conserved, and similar assignments can be made for the tau lepton and its neutrino.

In all cases it is important to keep in mind that lepton flavor conservation is not absolute, and that lepton flavor-violating reactions have a small probability of occurrence.

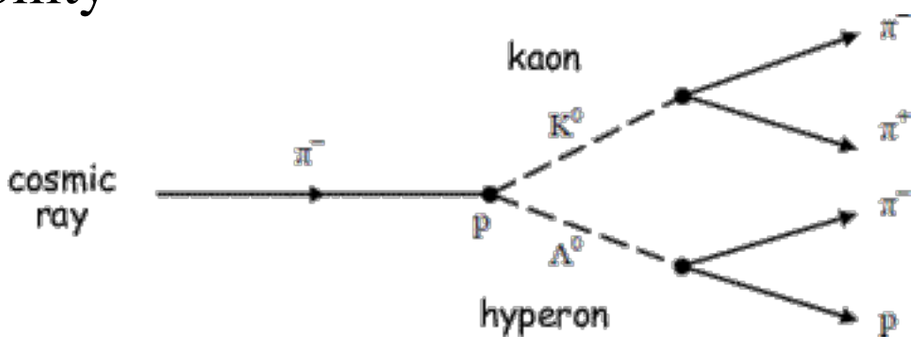
STRANGE PARTICLES AND STRANGENESS

Many particles discovered in the 1950s were produced by the nuclear interaction of pions with protons and neutrons in the atmosphere.

Three of these particles—namely, the kaon (K), lambda (Λ), and sigma (Σ) particles—exhibited unusual properties in production and decay and hence were called strange particles.

One unusual property of these particles is their production in pairs.

For example, when a pion collides with a proton, two neutral strange particles are produced with high probability



$$\pi^- + p \longrightarrow K^0 + \Lambda^0$$

On the other hand, the reaction $\pi^- + p \longrightarrow K^0 + n$ never occurred, even though no known conservation laws would have been violated and the energy of the pion was sufficient to initiate the reaction.

The second peculiar feature of strange particles is that although they are produced by the strong interaction at a high rate, they do not decay into particles that interact via the strong force at a very high rate.

Instead, they decay very slowly, which is characteristic of the weak interaction, as shown in the table.

Their half-lives are in the range 10^{-10} s to 10^{-8} s; most other particles that interact via the strong force have lifetimes on the order of 10^{-20} s and less.

To codify these unusual properties of strange particles, a law called conservation of strangeness was introduced with a new quantum number S, called **strangeness**.

The strangeness numbers for some particles are given in the table.

The production of strange particles in pairs is explained by assigning S = +1 to one of the particles and S = -1 to the other.

All nonstrange particles are assigned strangeness S = 0.

The **law of conservation of strangeness** states that *whenever a nuclear reaction or decay occurs, the sum of the strangeness numbers before the process must equal the sum of the strangeness numbers after the process*.

The slow decay of strange particles can be explained by assuming that the strong and electromagnetic interactions obey the law of conservation of strangeness but the weak interaction does not.

Because the decay reaction involves the loss of one strange particle, it violates strangeness conservation and hence proceeds slowly via the weak interaction.

the

the the the the the the

HOW ARE ELEMENTARY PARTICLES PRODUCED AND PARTICLE PROPERTIES MEASURED?

Examination of the bewildering array of entries in the earlier table leaves one yearning for firm ground.

It is natural to wonder about a particle that exists for 10^{-20} s and has a mass of 1192.5 MeV/c^2.

How is it possible to detect a particle that exists for only 10^{-20} s?

Furthermore, how can the mass be measured?

If a standard attribute of a particle is some type of permanence or stability, in what sense is a fleeting entity that exists for 10^{-20} s a particle?

Now we attempt to answer such questions and explain how elementary particles are produced and how their properties are measured.

Elementary particles, most of which are unstable and occur rarely naturally, are created abundantly in man-made collisions of high-energy particles with a suitable target.

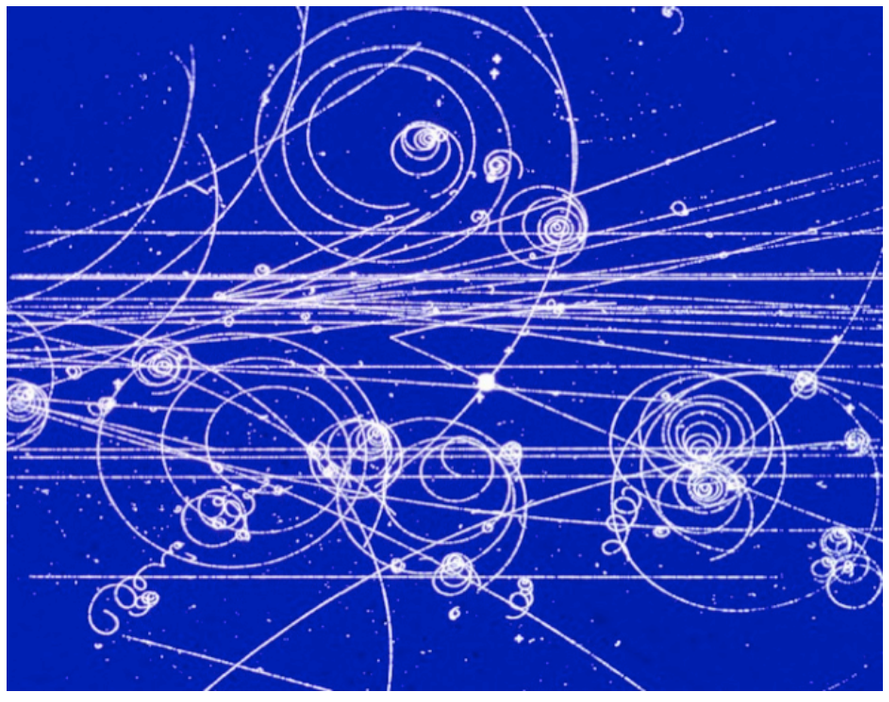
Since very high energy beams of incident particles are desirable, stable charged particles such as electrons or protons generally make up the incident beam, because it takes considerable time to accelerate particles to high energies with electromagnetic fields.

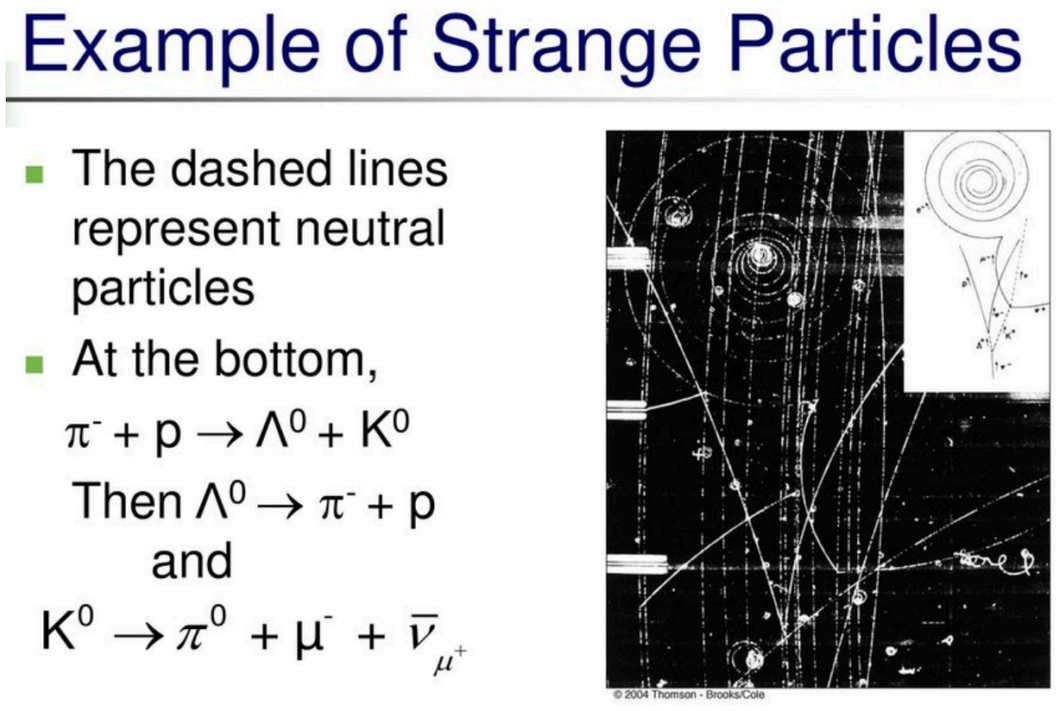
Similarly, targets must be simple and stable, and the simplest target, hydrogen, serves nicely as both a target and a detector.

In a liquid hydrogen bubble chamber, which is basically a large container filled with hydrogen near its boiling point, a charged particle traversing the chamber ionizes the atoms along its path, and the ionization causes a visible track of tiny bubbles.

The liquid hydrogen also serves as an efficient source of target protons, with a proton density sufficient to ensure many incident particle–target collisions within a reasonable time.

Figure shows a typical event in which a bubble chamber has served as both target and detector.





On the left is a generic bubble chamber image.,

On the right many parallel tracks of negative pions are visible entering the photograph from the bottom.

One of the pions has hit a stationary proton in the hydrogen and produced two strange particles, the Λ^0 and K^0 , according to the reaction

$$\pi^- + p \longrightarrow \Lambda^0 + K^0$$

Neither neutral strange particle leaves a track, but their subsequent decays into charged particles can be seen clearly, as indicated in the figure.

A magnetic field directed into the plane of the photograph causes the track of each charged particle to curve, and from the measured curvature the particle's charge and linear momentum can be determined.

If the mass and momentum of the incident particle are known, we can then usually calculate the product particle mass, kinetic energy, and speed from conservation of momentum and energy.

Finally, combining a product particle's speed with a measurable decay track length, we can calculate the product particle's lifetime.

The figure shows that sometimes one can use this lifetime technique even for a neutral particle, which leaves no track. As long as the particle speed and the start and finish of the missing track are known, one can infer the missing track length and find the lifetime of the neutral particle.

Resonance Particles

With clever experimental technique and much effort, decay track lengths as short as 1 micron (10^{-6} m) can be measured.

This means that lifetimes as short as 10^{-16} s can be measured with this technique in the case of high-energy particles traveling at about the speed of light.

We arrive at this result by assuming that a decaying particle travels 1 micron in the laboratory at a speed of 0.99c, yielding a laboratory lifetime of $\tau_{lab} = 10^{-6}$ m/0.99c $\approx 0.33 \times 10^{-14}$ s.

Now, relativity helps us.

Since the proper lifetime, as measured in the decaying particle's rest frame, is shorter than τ_{lab} by a factor of $\sqrt{1 - v^2/c^2}$, we can actually measure lifetimes of duration:

$$\tau_{proper} = \tau_{lab}\sqrt{1 - \frac{v^2}{c^2}} = (0.33 \times 10^{-14}\,\text{s})\sqrt{1 - \frac{(0.99c)^2}{c^2}} = 4 \times 10^{-16}\,\text{s}$$

Unfortunately, even with Einstein's help, we are several orders of magnitude away from minimum hadron lifetimes of 10^{-23} s with the best efforts of the track-length method.

How then, can we detect the presence of particles that exist for as short a time as 10^{-23} s?

As we shall see shortly, the masses, lifetimes, and very existence of these very short-lived particles, known as **resonance particles**, can be inferred from peaks (resonances) in the cross section versus energy plots describing their decay products.

Let's consider this in more detail by looking at the case of the resonance particle called the delta plus (Δ^+), which has a mass of 1231 MeV/c^2 and a lifetime of about 6 x 10^{-24} s.

The Δ^+ is produced in the reaction

$$e^- + p \longrightarrow e^- + \Delta^+$$

which is followed in 6 x 10^{-24} s by the decay

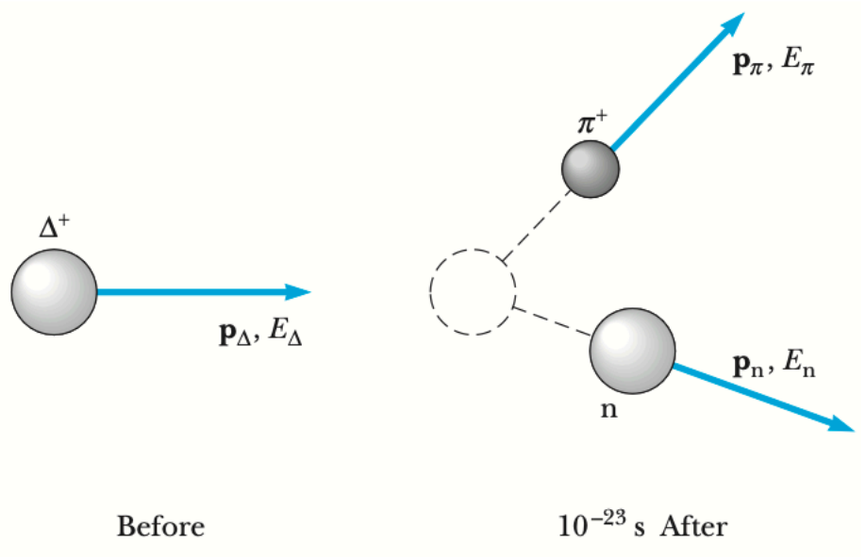
$$\Delta^+ \longrightarrow \pi^+ + n$$

Because the lifetime is so short, it leaves no measurable track, and it might seem impossible to distinguish the two reactions given above from the net direct reaction in which no Δ^+ is produced:

$$e^- + p \longrightarrow e^- + \pi^+ + n$$

In fact, we can tell whether a Δ^+ was formed by measuring the momentum and energy of the suspected decay products (pion and neutron) and using the conservation of momentum and energy to see whether these values combine to give a mass of 1231 MeV/c^2.

To understand this in detail, consider the decay of the shown in the figure below.



The decay of a Δ^+ particle into a positive pion and a neutron.

The energy and momentum of the must satisfy the equation

$$E_\Delta^2 = (p_\Delta c)^2 + (m_\Delta c^2)^2$$

so

$$m_\Delta c^2 = \sqrt{E_\Delta^2 - (p_\Delta c)^2}$$

where m_{Δ} is the Δ^+ mass.

Although we cannot directly measure E_{Δ} and $\mathbf{p}_{\Delta}$, since the delta particle leaves no track, we can measure the energies and momenta of the outgoing particles, E_{π}, $\mathbf{p}_{\pi}$, E_n, and $\mathbf{p}_n$.

Using conservation of momentum and energy, we can then find an expression for mc^2 in terms of these measured quantities.

Thus we have $E_{\Delta} = E_{\pi} + En$ and $\mathbf{p}_{\Delta} = \mathbf{p}_{\pi} + p_n$, and substituting into the equation,

$$m_{\Delta}c^2 = \sqrt{(E_{\pi} + E_n)^2 - (\mathbf{p}_{\pi} + \mathbf{p}_n)^2 c^2}$$

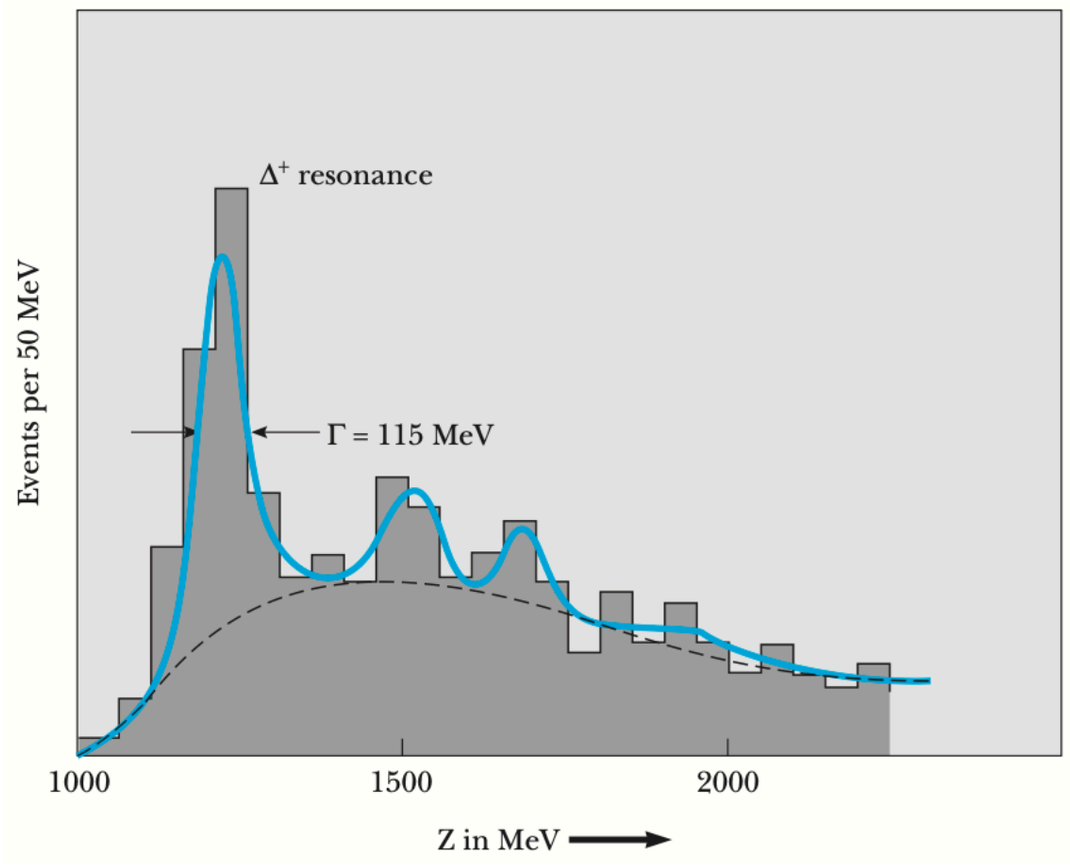
The equation holds for all events in which a particle actually formed and decayed.

That is, for many different measured values of E_{π}, $\mathbf{p}_{\pi}$, E_n, and $\mathbf{p}_n$ corresponding to many repeated experiments, we will always find the same value of the quantity $m_{\Delta}c^2 = 1231$ MeV within experimental uncertainty if the decay of Δ^+ is involved.

On the other hand, if no Δ^+ is involved and the direct reaction $e^- + p \rightarrow e^- + \pi^+ + n$ occurs, $\sqrt{(E_{\pi} + E_n)^2 - (\mathbf{p}_{\pi} + \mathbf{p}_n)^2 c^2}$ not equal 1231 MeV but will sweep over a broad range of values, some larger and some smaller than $m_{\Delta}c^2$, as the experiment is repeated.

The typical method for showing the existence of a resonance particle involves calculating the quantity $Z = \sqrt{(E_\pi + E_n)^2 - (\mathbf{p}_\pi + \mathbf{p}_n)^2 c^2}$ for a large number of events in which a π^+ and a neutron are produced and then drawing a histogram of those events.

By plotting the number of events with a given value of Z against this value of Z, we obtain a slowly varying curve with sharp peaks superimposed, the peaks showing the existence of resonance particles.



Experimental evidence for the existence of the Δ^+ particle. The sharp peak near 1230 MeV was produced by the events in which a Δ^+formed and promptly decayed into π^+ and neutron.

Figure shows such an experimental histogram for the particle.

The broad background (dashed curve) is produced by direct events in which no was created; the sharp peak near Z =1230 MeV, containing many events, was produced by all the events in which a formed and decayed into a pion and neutron with just the right energies and momenta to make up a delta particle.

Peaks corresponding to two other resonance particles of larger mass can also be seen in the figure.

Histograms las in the figure can tell us not only the mass of a short-lived particle but also the lifetime of the particle from the full width at half maximum of the resonance peak, denoted by in the figure.

Because Γ is twice the uncertainty in energy of the delta, ($\Gamma = 2\,\Delta$E), we can use the energy–time uncertainty relation, ΔE Δt $\approx \hbar/2$, to infer the lifetime of the delta, Δt:

$$\Delta t \approx \frac{\hbar}{2\,\Delta E} = \frac{\hbar}{\Gamma}$$

The measured width of $\Gamma = 115$ MeV leads to a value for the lifetime of the delta particle:

$$\Delta t \approx \frac{\hbar}{\Gamma} = \frac{6.6 \times 10^{-16}\ \text{eV}\cdot\text{s}}{115 \times 10^{6}\ \text{eV}} = 5.7 \times 10^{-24}\ \text{s}$$

In this incredibly short lifetime, a delta particle moving at the highest possible speed of c travels only 10^{-15} m, or about one nuclear diameter.

Energy Considerations in Particle Production

It is ironic that the highly sophisticated branch of physics known as elementary particle physics relies for its very existence on the most brutish of experimental methods: smashing an incident particle moving at relativistic speed into a stationary target particle and observing what pieces come flying out!

Yet this process is at the heart of experimental particle physics, and now we shall determine the threshold energy that is required for the production of new particles in a collision.

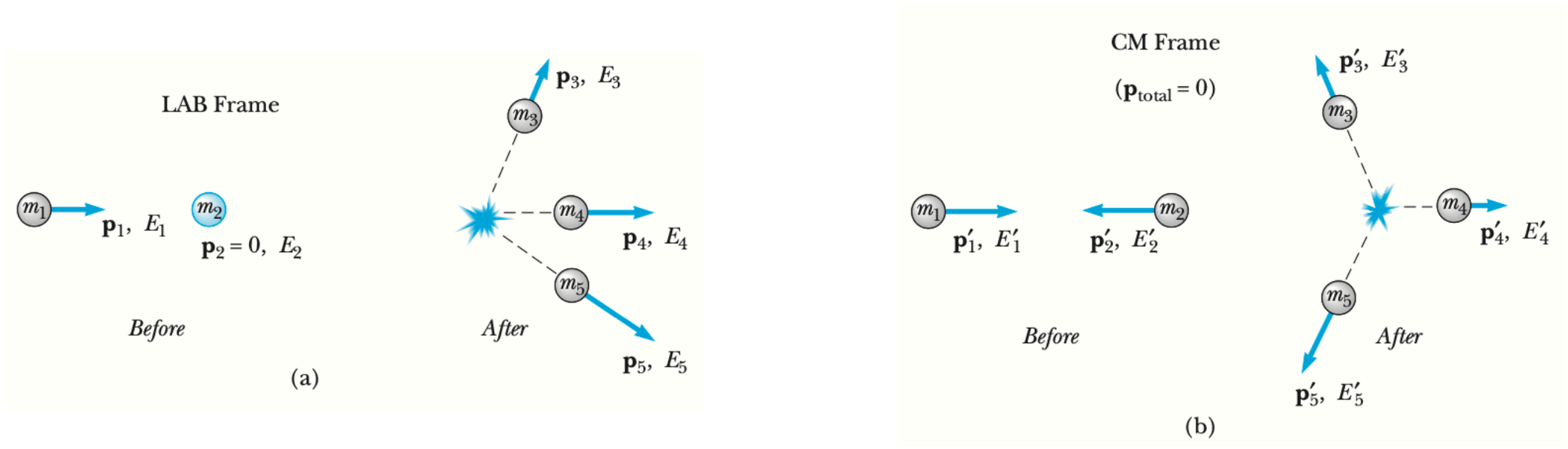
Since the energy needed to manufacture new particles comes from the kinetic energy of the incident particle, and incident particle energies are quite large (the Fermi National Laboratory produces 1000-GeV protons), we must use relativistic equations in the calculation of these threshold energies.

Consider the specific particle production process

$$m_1 + m_2 \longrightarrow m_3 + m_4 + m_5$$

Here m_1 is the mass of the incident particle, m_2 is the mass of the target particle that is at rest in the laboratory, and m_3, m_4, and m_5 are product-particle masses.

Figure part a shows such a particle reaction in the laboratory reference frame.



(a) The reaction $m_1 + m_2 \rightarrow m_3 + m_4 + m_5$ viewed in the laboratory frame, where m_1 is initially at rest and the energies are total relativistic energies (kinetic plus rest energy). (b) The same reaction viewed from the center-of-mass (CM) frame, in which the total momentum is always zero.

The energies shown in the figure — E_1, E_2, and so on — are total energies (kinetic energy rest energy) and the momenta are labeled $\mathbf{p}_1$, $\mathbf{p}_2$, and so on.

At first glance it might seem that the total initial energy of m_1 and m_2, $E_1 + E_2$, could be converted completely to rest energy of particles 3, 4, and 5.

However, as we have already seen in the calculation of nuclear reaction thresholds earlier, some of the initial energy must go into energy of motion of the product particles in order to conserve momentum in the collision.

This means that not all of the initial energy can go into creating new masses.

The issue is to find out how much of the initial energy can go toward creating masses m_3, m_4, and m_5.

That is, we wish to calculate the minimum or **threshold kinetic energy K_{th}** that m_1 must have in order to create particles with masses of m_3, m_4, and m_5 and also conserve momentum.

An important clue to finding K_{th} comes from understanding that if we could somehow arrange for a zero initial momentum, all of the initial energy could be converted into new particles.

What we can do is solve for K_{th} in the frame in which the initial momentum is zero—the center-of-mass frame— and then transform that result back into the frame in which the experiment is actually carried out, the laboratory frame.

Here's how it goes.

Figure above part b shows the same reaction as in part a but viewed from the center-of-mass (CM) frame, in which, by definition, the total momentum is always zero.

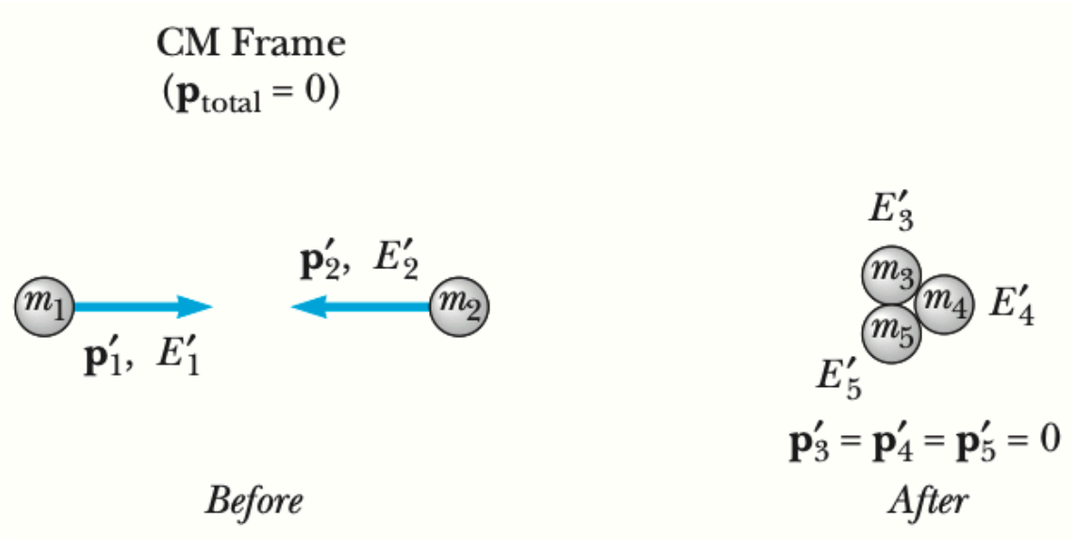
Thus, in part b, m_1 and m_2 have equal and opposite momenta, and the vector sum of the momenta of m_3, m_4, and m_5 is zero.

Actually, part b shows the case of an incident particle with more than threshold energy.

A moment's reflection reveals that when m_1 has the minimum, or threshold, energy in the CM frame, all of the initial energy, $E'_1 + E'_2$, should be converted to the masses of particles 3, 4, and 5.

This occurs when the product particles are created at rest in the CM frame, which is now possible since the total momentum is zero in the CM frame.

The figure below shows the reaction $m_1 + m_2 \longrightarrow m_3 + m_4 + m_5$ in the CM frame when m_1 has the threshold kinetic energy and m_3, m_4, and m_5 are created at rest.



The reaction $m_1 + m_2 \longrightarrow m_3 + m_4 + m_5$ in the CM frame when m_1 has the threshold kinetic energy required to produce m_3, m_4, and m_5 . For the threshold condition, m_3, m_4, and m_5 are created at rest.

To calculate a numerical expression for K_{th}, we make use of the invariant quantity $E^2 - p^2c^2$, introduced in when we discussed relativity.

Recall that $E^2 - p^2c^2$ is called an invariant because it has the same numerical value for a system of particles in any inertial frame.

Applying the invariance of $E^2 - p^2c^2$ to the CM and laboratory frames, we have, for the period before the collision,

$$E_{\text{CM}}^2 - p_{\text{CM}}^2 c^2 = E_{\text{lab}}^2 - p_{\text{lab}}^2 c^2$$

Here, E_{CM} and E_{lab} stand for the sum of the energies of particles m_1 and m_2 in the CM and laboratory frames, respectively.

Likewise, p_{CM} and p_{lab} stand for the sum of the momenta of particles m_1 and m_2 in the CM and laboratory frames.

Because $p_{CM} = 0$, $E_{lab} = E_1 + m_2c^2$, and $p_{lab} = p_1$, Equation 15.11 becomes

$$E_{\text{CM before}}^2 = (E_1 + m_2 c^2)^2 - p_1^2 c^2$$

We can eliminate p_1 from the above equation by using $E_1^2 = p_1^2c^2 + m_1^2c^3$ to obtain

$$E_{\text{CM before}}^2 = 2E_1 m_2 c^2 + (m_2^2 + m_1^2) c^4$$

Observe that this equation is in a useful form, since we can readily solve for E_1 or K_{th} ($E_1 = K_{th} + m_1c^2$) in terms of all the masses if we can find an expression for E_{CM} before in terms of the product masses m_3, m_4, and m_5.

We can find such an expression by conserving relativistic energy in the CM frame:

$$E_{\text{CM before}} = E_{\text{CM after}} = E_3' + E_4' + E_5' = (m_3 + m_4 + m_5)c^2$$

Substituting this expression above and using $E_1 = K_{th}\, m_1c^2$, after a bit of algebra we obtain our final result:

$$K_{\text{th}} = \frac{(m_3 + m_4 + m_5)^2c^2 - (m_1 + m_2)^2c^2}{2m_2}$$

This result gives the threshold kinetic energy of an incident particle m_1 colliding with a stationary target m_2 required to produce three product particles of mass m_3, m_4, and m_5.

For more than three product particles, inspection of the derivation shows the result can be generalized as

$$K_{\text{th}} = \frac{(m_3 + m_4 + m_5 + m_6 + \cdots)^2c^2 - (m_1 + m_2)^2c^2}{2m_2}$$

THE EIGHTFOLD WAY

One of the tools scientists use is the detection of patterns in data, patterns that contribute to our understanding of nature.

One of the best examples of the use of this tool is the development of the periodic table, which provides a fundamental understanding of the chemical behavior of the elements.

The periodic table explains how more than 100 elements can be formed from three particles—the electron, the proton, and the neutron.

The table of nuclides contains hundreds of nuclides, but all can be built from protons and neutrons.

The number of observed particles and resonances observed by particle physicists is also in the hundreds.

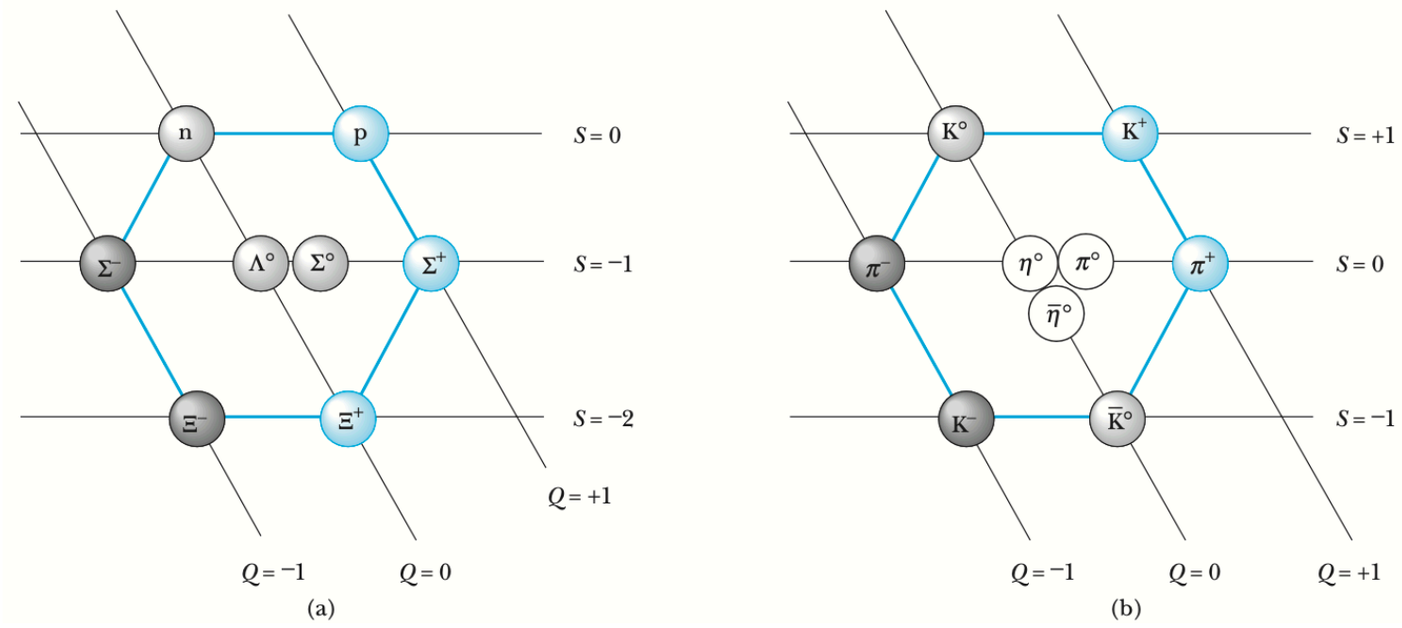
Is it possible that a small number of entities exist from which all of these can be built?

Taking a hint from the success of the periodic table and the table of nuclides, let us explore the historical search for patterns among the particles.

Many classification schemes have been proposed for grouping particles into families.

Consider, for instance, the baryons listed in the earlier table that have spins 1/2 :
p, n, Λ^0, Σ^+, Σ^0, Σ^-, Ξ^0, and Ξ^- .

If we plot strangeness versus charge for these baryons using a sloping coordinate system, as in the figure part a below, we observe a fascinating pattern: Six of the baryons form a hexagon, and the remaining two are at the hexagon's center.



(a) The hexagonal eightfold-way pattern for the eight spin-1/2 baryons. This strangeness-versus-charge plot uses a sloping axis for the charge number Q but a horizontal axis for the strangeness (S) values. (b) The eightfold-way pattern for the nine spin-0 mesons.

As a second example, consider the following nine spin-zero mesons listed in the earlier table: , π^+, π^0, π^-, K^+, K^0, K^-, η, η' , and the antiparticle $\overline{K}^0$.

Figure above part b is a plot of strangeness versus charge for this family.

Again, a hexagonal pattern emerges.

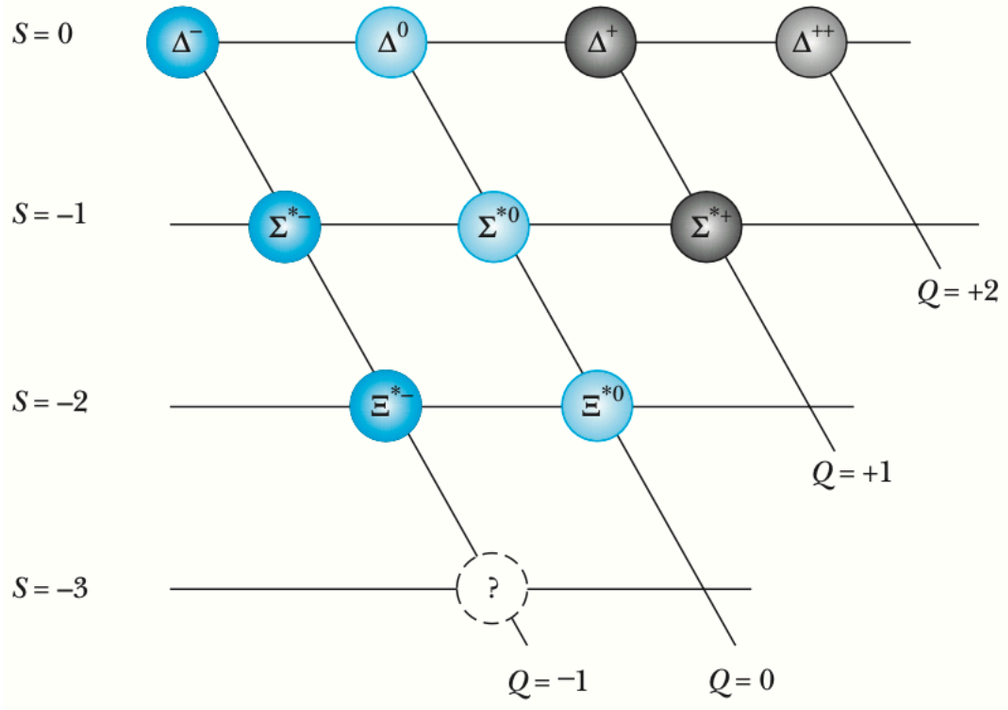
In this case, each particle on the perimeter of the hexagon lies opposite its antiparticle, and the remaining three (which form their own antiparticles) are at the center of the hexagon.

These and related symmetric patterns were developed independently in 1961 by Murray Gell-Mann and Yuval Ne'eman (b. 1925).

Gell-Mann called the patterns the **eightfold way**, after the eightfold path to nirvana in Buddhism.

Groups of baryons and mesons can be displayed in many other symmetric patterns within the framework of the eightfold way.

For example, the family of spin-3/2 baryons known in 1961 contained nine particles arranged in a pattern like that of the pins in a bowling alley, as in the figure below.



The pattern for the higher-mass, spin- 3/2 baryons known at the time the pattern was proposed. The three and two Σ^* particles are excited states of the Ξ^* corresponding spin- 1/2 particles in the earlier figure. These excited states have higher mass and spin 3/2. The absence of a particle in the bottom position was evidence of a new particle yet to be discovered, the Ω^-..

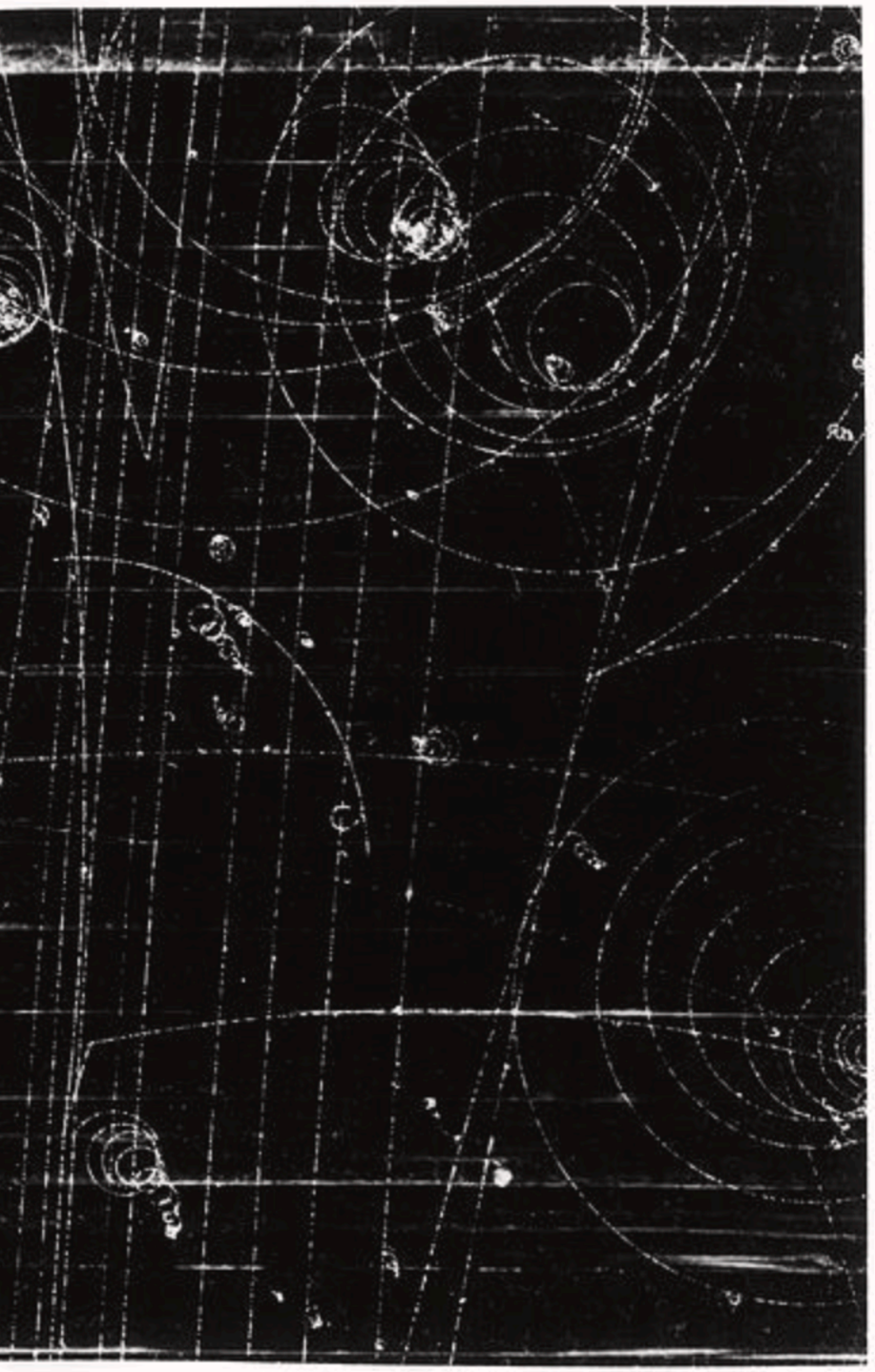
When this pattern was proposed, an empty spot occurred in it (at the bottom position), corresponding to a particle that had never been observed.

Gell-Mann predicted that the missing particle, which he called the omega minus(Ω^-.), should have spin 3/2, charge -1, strangeness -3, and rest energy of approximately 1680 MeV.

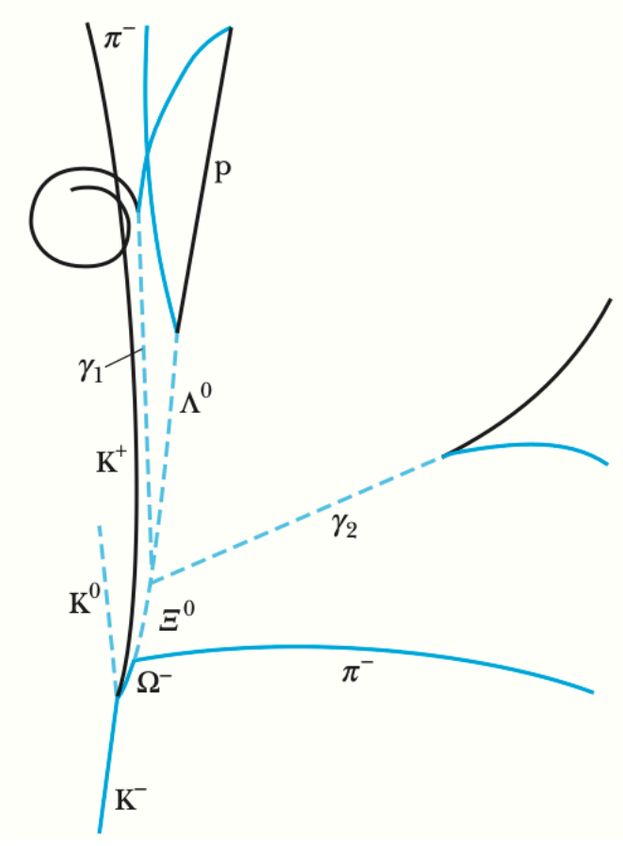
Shortly thereafter, in 1964, scientists at the Brookhaven National Laboratory found the missing particle through careful analyses of bubble-chamber photographs and confirmed all its predicted properties (figure below).

The prediction of the missing particle in the eightfold way has much in common with the prediction of missing elements in the periodic table.

Whenever a vacancy occurs in an organized pattern of information, experimentalists have a guide for their investigations.



Discovery of the Ω^-. particle. The photograph on the left shows the original bubble-chamber tracks. The drawing on the right isolates the tracks of the important events. The K^- particle at the bottom collides with a proton to produce the first detected particle plus a K^0 and a K^+



the

QUARKS

As we have noted, leptons appear to be truly elementary particles because they have no measurable size or internal structure, are limited in number, and do not seem to break down into smaller constituents

The existence of the eightfold-way patterns suggests that baryons and mesons—in other words, hadrons—have a more elemental substructure.

Furthermore, we know that hadrons decay into other hadrons and are many in number.

The table below lists only the hadrons that are stable against hadronic decay; hundreds of others have been discovered.

These facts strongly suggest that hadrons cannot be truly elementary.

Table Properties of Quarks and Antiquarks

Quarks

Name	Symbol	Spin	Charge	Baryon Number	Strangeness	Charm	Bottomness	Topness
Up	u	$\frac{1}{2}$	$+\frac{2}{3}e$	$\frac{1}{3}$	0	0	0	0
Down	d	$\frac{1}{2}$	$-\frac{1}{3}e$	$\frac{1}{3}$	0	0	0	0
Strange	s	$\frac{1}{2}$	$-\frac{1}{3}e$	$\frac{1}{3}$	−1	0	0	0
Charmed	c	$\frac{1}{2}$	$+\frac{2}{3}e$	$\frac{1}{3}$	0	+1	0	0
Bottom	b	$\frac{1}{2}$	$-\frac{1}{3}e$	$\frac{1}{3}$	0	0	+1	0
Top	t	$\frac{1}{2}$	$+\frac{2}{3}e$	$\frac{1}{3}$	0	0	0	+1

Antiquarks

Name	Symbol	Spin	Charge	Baryon Number	Strangeness	Charm	Bottomness	Topness
Anti-up	$\bar{u}$	$\frac{1}{2}$	$-\frac{2}{3}e$	$-\frac{1}{3}$	0	0	0	0
Anti-down	$\bar{d}$	$\frac{1}{2}$	$+\frac{1}{3}e$	$-\frac{1}{3}$	0	0	0	0
Anti-strange	$\bar{s}$	$\frac{1}{2}$	$+\frac{1}{3}e$	$-\frac{1}{3}$	+1	0	0	0
Anti-charmed	$\bar{c}$	$\frac{1}{2}$	$-\frac{2}{3}e$	$-\frac{1}{3}$	0	−1	0	0
Anti-bottom	$\bar{b}$	$\frac{1}{2}$	$+\frac{1}{3}e$	$-\frac{1}{3}$	0	0	−1	0
Anti-top	$\bar{t}$	$\frac{1}{2}$	$-\frac{2}{3}e$	$-\frac{1}{3}$	0	0	0	−1

Now we show that the complexity of hadrons can be explained by a simpler substructure.

The Original Quark Model

In 1963 Gell-Mann and George Zweig independently proposed that hadrons have a more elemental substructure.

According to their model, all hadrons are composite systems of two or three fundamental constituents called quarks.

Gell-Mann borrowed the word quark from the passage “Three quarks for Muster Mark” in James Joyce’s Finnegan’s Wake.

In the original quark model, there were three types of quarks designated by the symbols u, d, and s.

These were given the arbitrary names **up**, **down**, and *sideways* or, now more commonly, **strange**.

An unusual property of quarks is that they had to be assigned fractional electronic charges.

The u, d, and s quarks have charges of 2e/3, e/3, and e/3, respectively, where e is the charge of a positron.

Each quark has a baryon number of 1/3 and a spin of 1/2.

The u and d quarks have strangeness 0, and the s quark has strangeness -1.

Associated with each quark is an anti-quark of opposite charge, baryon number, and strangeness.

Earlier table gives other properties of quarks and antiquarks.

The compositions of all hadrons that were known when Gell-Mann and Zweig presented their models could be completely specified by three simple rules:

- A meson consists of one quark and one antiquark, which gives it a baryon number of 0, as required. Mesons are thus quark–antiquark combinations (quarkonium) bound together by a swarm of gluons, which are the field particles that transmit the strong force between quarks.
- A baryon consists of three quarks and is a sort of quark "molecule" held together by gluons.
- An antibaryon consists of three antiquarks.

Table lists the quark compositions of several mesons and baryons.

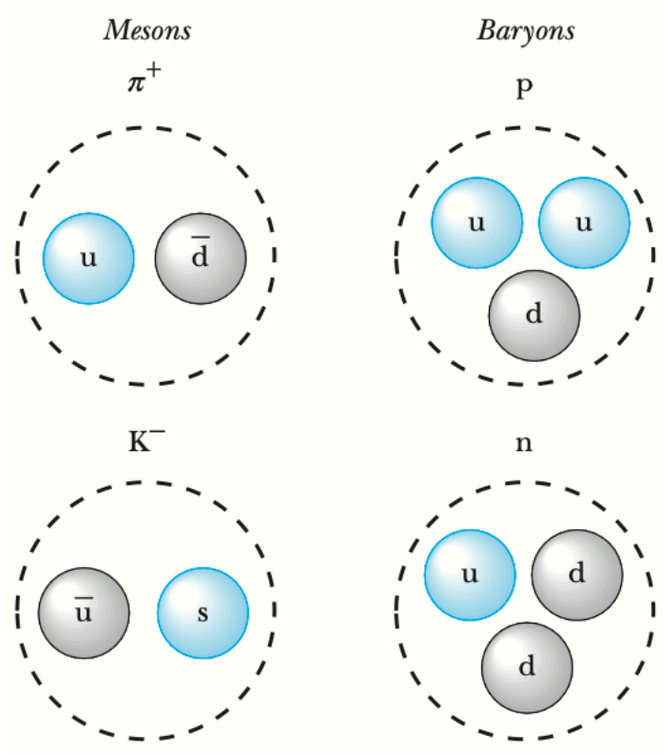
Note that just two of the quarks, u and d, are contained in all hadrons encountered in ordinary matter (protons and neutrons).

The third quark, s, is needed only to construct strange particles with a strangeness number of either 1+ or -1.

Figure below is a pictorial representation of the quark compositions of several particles.

Table Quark Compositions of Several Hadrons

Particle	Quark Composition
Mesons	
π^+	$u\bar{d}$
π^-	$\bar{u}d$
K^+	$u\bar{s}$
K^-	$\bar{u}s$
K^0	$d\bar{s}$
Baryons	
p	uud
n	udd
Λ^0	uds
Σ^+	uus
Σ^0	uds
Σ^-	dds
Ξ^0	uss
Ξ^-	dss
Ω^-	sss



Quark compositions of two mesons and two baryons. Note that each meson on the left contains two quarks, while each baryon on the right contains three quarks

Charm and Other Developments

Although the original quark model was highly successful in classifying particles into families, some discrepancies occurred between predictions of the model and certain experimental decay rates.

Consequently, several physicists proposed a fourth quark in 1967.

The fourth quark, denoted by c, was given a property **charm**.

A charmed quark would have charge 2e/3, but its charm would distinguish it from the other three quarks.

The new quark would have a charm of C = +1, its antiquark would have a charm of C = -1, and all other quarks would have C = 0, as indicated in the earlier table.

Charm, like strangeness, would be conserved in strong and electromagnetic interactions but not in weak interactions.

Evidence that the charmed quark exists began to accumulate in 1974 when a new heavy particle called the J/Ψ particle (or simply Ψ) was discovered independently by two groups, one led by Burton Richter at the Stanford Linear Accelerator (SLAC) and the other led by Samuel Ting at the Brookhaven National Laboratory.

Richter and Ting were awarded the Nobel prize in 1976 for this work.

Although massive, the J/Ψ particle was not a three-quark baryon but had the properties of a combination of a charmed quark and its antiquark $(c\bar{c})$.

It was much more massive (~3100 MeV/c^2) than the other known mesons, and its lifetime was much longer than those that decay via the strong force.

Soon, related charmed mesons were discovered that corresponded to such quark combinations as $\bar{c}d$ and $c\bar{d}$, all of which have large masses and long lifetimes.

The existence of these new mesons provided firm evidence for the fourth quark flavor.

In 1975 researchers at Stanford University reported strong evidence for the tau (τ) lepton, with a mass of 1784 MeV/c^2.

This was the fifth type of lepton to be discovered, which led physicists to propose that two new quarks, **top** (t) and **bottom** (b), might exist. (Some physicists prefer the designators truth and beauty.)

To distinguish these quarks from the old ones, quantum numbers called **topness** and **bottomness** (with allowed values +1, 0, -1) were assigned to all quarks and antiquarks (see earlier table).

In 1977 researchers at the Fermi National Laboratory, under the direction of Leon Lederman, reported the discovery of a very massive new meson, Y, whose composition is considered to be $b\bar{b}$.

In March 1995 researchers at the Fermi National Laboratory announced the discovery of the top quark (supposedly the last of the quarks), with a mass of 173 GeV/c^2.

The researchers identified t pairs from their decay into Wt bosons and bottom quarks and the subsequent decay of those particles into leptons, neutrinos, and hadrons.

The original $t\bar{t}$ pairs were produced by the collision of 0.9 TeV protons with 0.9 TeV antiprotons.

You are probably wondering whether such discoveries will ever end.

How many “building blocks” of matter really exist?

At the present, physicists believe that the fundamental particles in nature are 12 fermions, including 6 quarks and 6 leptons (together with their antiparticles) and the field particles listed in earlier table.

The table below lists properties of these 12 fermions as well as properties of the field particles excepting the graviton.

Table Properties of the Fundamental Point Particles

Particle		Approximate Rest Mass in Terms of Proton Mass[a]	Interactions Experienced[b]	Interaction Mediated	Electric Charge in Units of $\lvert e \rvert$	Color Charge
Leptons	e^-, μ, τ	$\frac{1}{1836}$, $\frac{1}{9}$, 1.9	EM, W	None	-1	No
	ν_e, ν_μ, ν_τ	$<3 \times 10^{-9}$, $<4 \times 10^{-9}$, $<9 \times 10^{-9}$	W	None	0	No
Quarks	u, c, t	$\frac{1}{2.8}$, 1.6, 185	EM, W, S	None	$+\frac{2}{3}$	Yes
	d, s, b	$\frac{1}{2.8}$, $\frac{1}{1.7}$, 5.3	EM, W, S	None	$-\frac{1}{3}$	Yes
All are fermions with spin $\frac{1}{2}$.						
Field Particles						
Photon	γ	0	EM	EM	0	No
Intermediate bosons	$W^\pm$	86	W, EM	W	± 1	No
	Z^0	97	W	W	0	No
Gluons	g	0	S	S	0	Yes
All are bosons with spin 1.						

[a]For quarks these are inferred values because free quarks have never been observed.

[b]EM = electromagnetic, W = weak, S = strong

Despite extensive experimental efforts, no isolated quark has ever been observed.

Physicists now believe that quarks are permanently confined inside hadrons because of an exceptionally strong force that prevents them from escaping.

This force, called the strong or “color” force, increases with separation distance, eventually approaching a constant value—implying a potential energy that grows linearly with quark separation and the requirement of infinite energy to produce two truly free quarks

COLORED QUARKS, OR QUANTUM CHROMODYNAMICS

Shortly after the concept of quarks was proposed, scientists recognized that certain particles had quark compositions that violated the Pauli exclusion principle.

Because quarks are fermions with spins of 1/2, they are expected to follow the exclusion principle.

One example of a particle that violates the exclusion principle is the Ω^- (s s s) baryon, which contains three s quarks with parallel spins, giving it a total spin of 3/2.

Other examples of baryons that have identical quarks with parallel spins are the Δ^{++}(u u u) and the Δ^-(d d d).

To resolve this problem, Moo-Young Han and Yoichiro Nambu suggested in 1965 that quarks possess a new property, called **color** or **color charge**.

This property is similar in many respects to electric charge except that it occurs in three varieties, called red, green, and blue.

The antiquarks have the colors antired, antigreen, and antiblue.

To satisfy the exclusion principle, all three quarks in a baryon must have *different* colors.

Just as a combination of actual colors of light can produce the neutral color white, a combination of three quarks with different colors is said to be white, or colorless.

A meson consists of a quark of one color and an antiquark of the corresponding anticolor.

The result is that both baryons and mesons are always colorless (or white).

Although the concept of color in the quark model was originally conceived to satisfy the exclusion principle, it also provided an improved theory for explaining certain experimental results.

For example, the modified theory correctly predicts the lifetime of the π^0 meson.

The general theory of how quarks interact with each other is called **quantum chromodynamics** (QCD), to parallel quantum electrodynamics, the theory of interaction between electric charges.

In QCD the quark is said to carry a color charge, in analogy with electric charge, and the **color charge** is responsible for the strong force between quarks.

The strong force between quarks is often called the **color force**.

The color force between quarks is analogous to the electric force between charges; like colors repel and opposite colors attract.

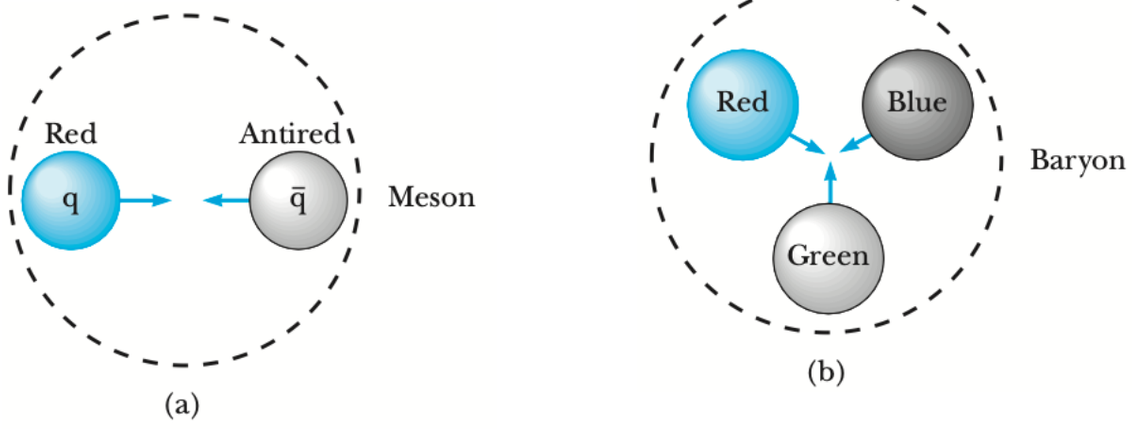
Therefore, two red quarks repel each other, but a red quark is attracted to an antired quark.

The attraction between quarks of opposite color to form a meson ($q\bar{q}$) is indicated in the figure below part a.

Differently colored quarks also attract each other, but with less intensity than the oppositely colored quark and antiquark.

For example, the red, blue, and green quarks in a cluster all attract each other to form baryons, as indicated in the figure below part b.

Thus every baryon contains three quarks of three different colors.



(a) A red quark is attracted to an antired quark. This forms a meson whose quark structure is ($q\bar{q}$)
(b) Three differently colored quarks attract each other to form a baryon.

As stated earlier, the strong force between quarks is carried by massless particles that travel at the speed of light called **gluons** (g).

According to QCD, there are eight gluons, *all carrying two color charges, a color and an anticolor*, such as a "blue-antired" gluon.

When a quark emits or absorbs a gluon, its color changes.

For example, a blue quark that emits a blue-antired gluon becomes a red quark, and a red quark that absorbs this gluon becomes a blue quark.

Because gluons carry color charge, they should clump together to form entities called glue-balls, but no glue-ball particles have yet been experimentally detected.

Experimental Evidence for Quarks

As already mentioned, an isolated quark has never been observed and probably never will be, according to QCD.

This leaves us with the question of how quarks can be detected and what experimental evidence we have for the existence of quarks.

Experiments at the Stanford Linear Accelerator (SLAC) in the late 1960s established two important results concerning quarks.

(1) The recoils of high-energy electrons scattered from protons could be completely modeled by assuming three point-like quarks within the proton.

(2) Although gluons were not directly detected, it was found that the incident electrons (which carry no color charge) interacted with only about half the mass making up the proton.

This is exactly what is expected if approximately half the material in a proton consists of gluons that are electrically neutral and do not interact electromagnetically with the electron (see earlier table).

It is interesting to consider what happens to a quark–antiquark pair ($q\bar{q}$) formed in a high-energy collision as the pair attempts to separate.

Because the color force is so strong, all quarks must form colorless baryons (three quarks) or mesons (two quarks).

The process goes like this.

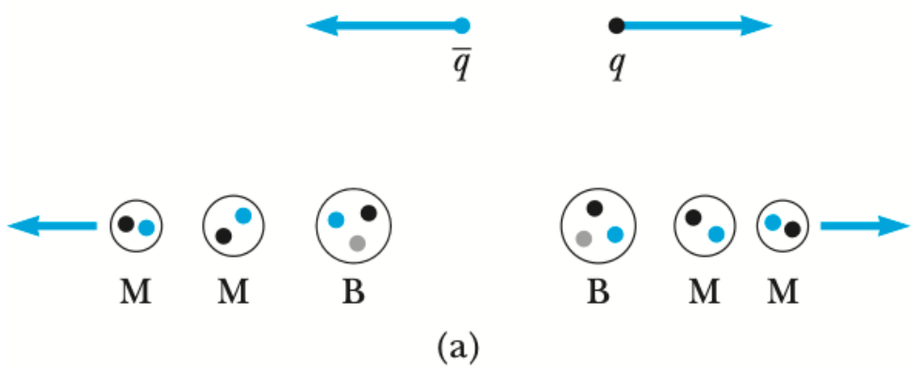
As two quarks separate, the increase in energy of the color field between them is so large that additional quark–antiquark pairs are created in streams (or jets) following the original ($q\bar{q}$) pair.

When all the available energy is used up, the quarks cluster into color-neutral or colorless combinations, which continue to separate indefinitely as two jets of hadrons, as in the figure below part a.

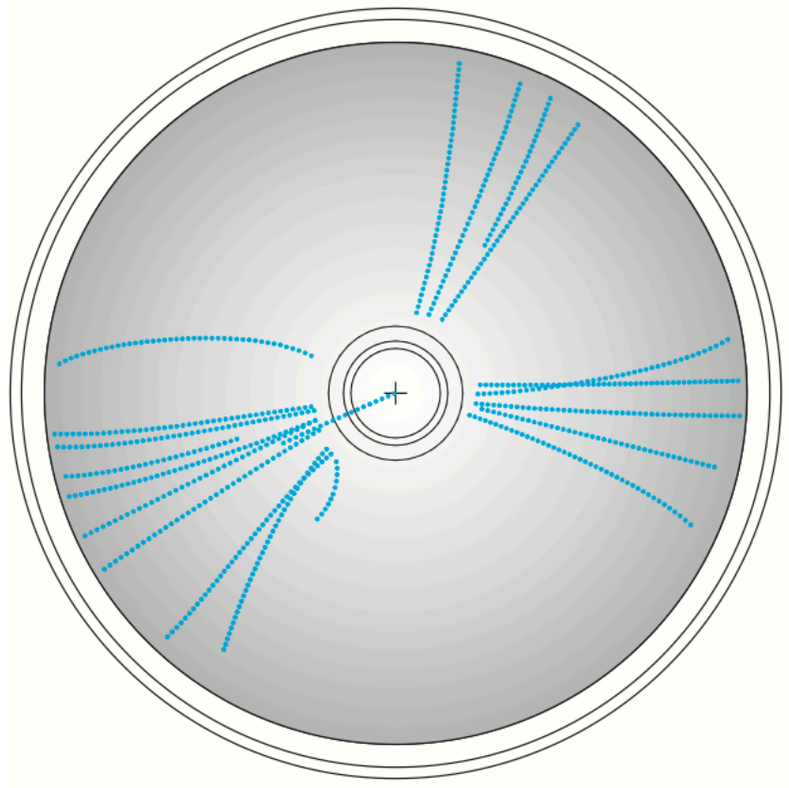
The figure below part b shows actual experimental data corresponding to three jets of hadrons emerging from a collision in which a quark and antiquark were produced and a gluon was radiated.

Because gluons carry color charge, they, like quarks, must also appear as a jet of hadrons.

Thus *experimental signatures of quarks and gluons are narrow "jets" of hadrons*.



(a)


(b)

(a) As the quark–antiquark pair separates, the potential energy in the color field is transformed into additional $q\bar{q}$ pairs, which quickly condense into colorless mesons (M) and baryons (B). These separate as jets of hadrons. (b) Three jets of hadrons produced by a quark, an antiquark, and a gluon. This figure is from the JADE detector at the German laboratory DESY in Hamburg.

Explanation of Nuclear Force in Terms of Quarks

Although the color force between two color-neutral hadrons is negligible at large separations, the strong color force between their constituent quarks does not exactly cancel at small separations.

This residual strong force is, in fact, the short-range nuclear force that binds protons and neutrons to form nuclei. Earlier, we explained the nuclear interaction of a proton with a neutron using Yukawa's early theory of pion exchange.

According to QCD, a more basic explanation of nuclear force can be given in terms of quarks and gluons, as shown by contrasting Feynman diagrams of the same process in the next figure.

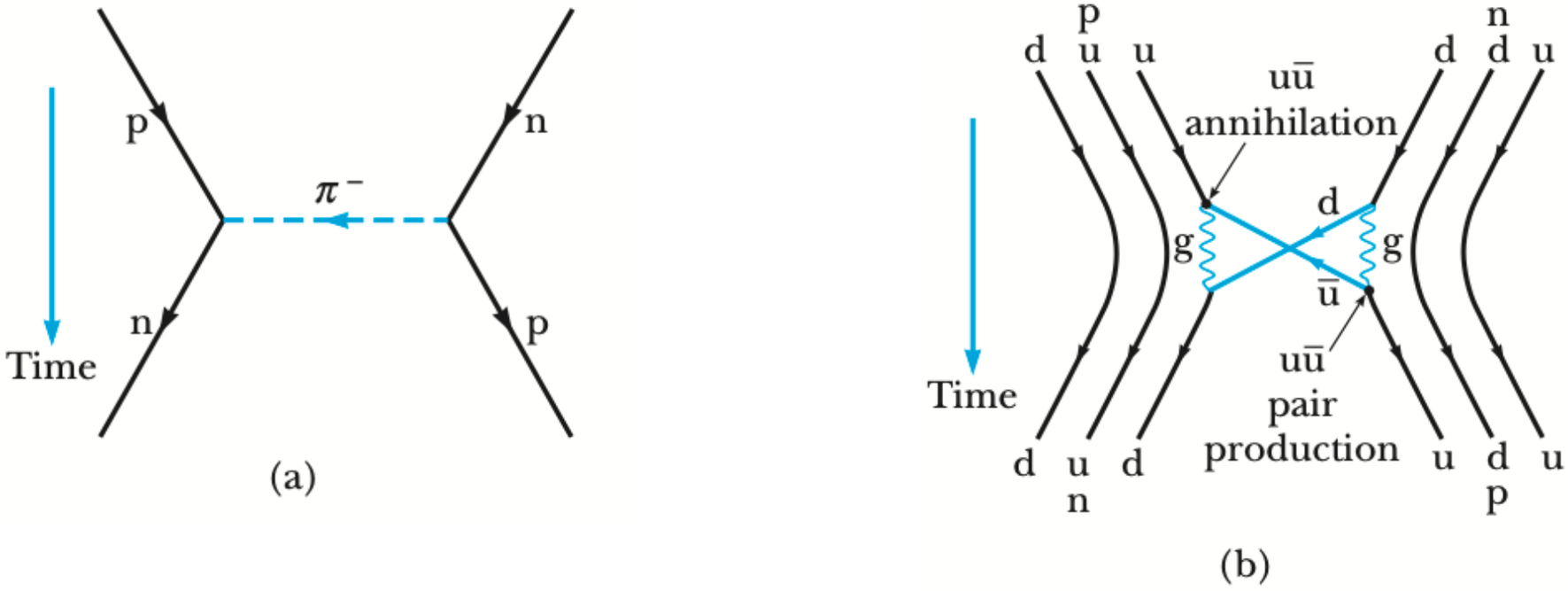
Each quark within the neutron and proton is continually emitting and absorbing virtual gluons and creating and annihilating virtual ($q\bar{q}$) pairs.

When the neutron and proton approach within 1 fm of each other, these virtual gluons and quarks can be exchanged between the two nucleons, and such exchanges produce the nuclear force.

Part b of the figure depicts one likely possibility or contribution to the general process shown in part a.

A down quark emits a virtual gluon (represented by a wavy line, g), which creates, a little later, a $u\bar{u}$ pair.

Both the recoiling d quark and the $\bar{u}$ are transmitted to the proton, where the $\bar{u}$ annihilates a proton u quark (with the creation of a gluon) and the d is captured.



(a) A nuclear interaction between a proton and a neutron explained in terms of Yukawa's pion exchange model.

(b) The same interaction as in (a), explained in terms of quarks and gluons. Note that the exchanged $\bar{u}d$ quark pair makes up a π^- meson.

ELECTROWEAK THEORY AND THE STANDARD MODEL

Recall that the weak interaction is an extremely short-range force with an interaction distance of approximately 10^{-18} m.

Such a short-range interaction implies that the quantized particles that carry the weak field (the spin-1 W^+, W^-, and Z^0 bosons) are quite massive, as is indeed the case (see earlier table).

These bosons are especially amazing when we realize they are structureless, point-like particles as massive as krypton atoms!

As mentioned earlier, the weak interaction is responsible for neutron decay and the beta decay of other, heavier baryons.

More important, the weak interaction is responsible for the decay of the c, s, b, and t quarks into lighter, more stable u and d quarks as well as the decay of the massive and leptons into (lighter) electrons.

Thus, the weak interaction is very important because it governs the stability of the basic matter particles.

A mysterious feature of the weak interaction is its lack of symmetry, especially in comparison with the high degree of symmetry shown by the strong, electromagnetic, and gravitational interactions.

For example, the weak interaction, unlike the strong interaction, is not symmetric under mirror reflection or charge exchange.

Mirror reflection means that all the quantities in a given particle reaction are exchanged as in a mirror reflection—left for right, an inward motion toward the mirror for an outward motion.

Charge exchange means that all the electric charges in a particle reaction are converted to their opposites —all positives to negatives and vice versa.

When we say that the weak interaction is not symmetric, we mean that the reaction with all quantities changed occurs less frequently than the direct reaction.

For example, the decay of the K^0, which is governed by the weak interaction, is not symmetric under charge exchange, since

$$K^0 \longrightarrow \pi^- + e^+ + \nu_e$$

occurs much more frequently than

$$K^0 \longrightarrow \pi^+ + e^- + \bar{\nu}_e$$

In 1979 Sheldon Glashow, Abdus Salam, and Steven Weinberg won a Nobel prize for developing a theory that unifies the electromagnetic and weak interactions.

This **electroweak theory** postulates that the weak and electromagnetic interactions have the same strength at very high particle energies.

Thus the two interactions are viewed as two different manifestations of a single, unifying electroweak interaction.

The photon and the three massive bosons ($W^{\pm}$ and Z^0) play key roles in the electroweak theory.

Perhaps the most spectacular of the theory's many concrete predictions was the prediction of the masses of the W and Z particles at about 80 GeV/c^2 and 91 GeV/c^2, respectively

The 1984 Nobel prize was awarded to Carlo Rubbia and Simon van der Meer for their work leading to the discovery of these particles with just those masses at CERN.

The combination of the electroweak theory and QCD for the strong interaction is called the **Standard Model** in high-energy physics.

It includes almost all the constituents of matter—six leptons, six quarks, and three forces and their field particles—but not the gravitational force at this time.

The Standard Model does not answer all questions.

It requires as input parameters over a dozen measured numbers such as the lepton and quark masses.

Further, it does not explain the mysterious ratios of these masses, why there are only six types of leptons, or why there are six types of both leptons and quarks.

A specific question regarding field particles is why the photon has no mass while the W and Z bosons do have mass.

Because of this difference, the electromagnetic and weak forces are quite distinct at low energies but become similar, or symmetric, at very high energies, where the rest energies of the W and Z bosons are insignificant fractions of their total energies.

This behavior in the transition from high to low energies is called **symmetry breaking**.

In 1964 Peter Higgs (b. 1929, Scottish physicist) introduced a mechanism for electroweak symmetry breaking by proposing a new field, called the **Higgs field**, which permeates all of space and gives particles their mass.

Roughly, the Higgs field may be viewed as causing a kind of drag force on particles as they interact with it, giving particles their characteristic inertia.

As with all other classical fields, quantization of the Higgs field results in a force-carrying particle called the **Higgs boson**.

The Standard Model, including the Higgs mechanism, provides a logically consistent explanation of the massive nature of the W and Z bosons.

The Higgs boson has been observed experimentally using hte Large Hadron Collider with all the correct properties.

BEYOND(??) THE STANDARD MODEL

The Standard Model works!

It explains every known experiment that has been carried out since the 1970s.

That, of course, has not stopped theorists from attempted grander theories.

Some of the names are **grand unification theory**, or **GUT** , **Supersymmetry (SUSY)** , **String Theory** , **M-Theory** and so on.

None of the new ideas have made a single prediction that is different than the Standard Model.

Most have made incorrect predictions - the Standard model has none!

String Theory makes no predictions!!!!

As my thesis advisor Professor Hans Bethe said in the 1970s, there is nothing beyond the Standard Model until you get the energies like 10^{19} GeV (LHC is 10^{4} GeV) - he called it the great desert!

We are now done!